Deep Learning for Seed Vigor Assessment: Analysis of U-Net Family Architectures in Soybean Seedling Segmentation

José Luciano M. Neto[2], Emannuel Diego G. de Freitas[1,2], Pedro Henrique dos S. e Silva[1], Haynna F. Abud[3], Danielo G, Gomes[1]

[1]Departamento de Engenharia de Teleinformática, Universidade Federal do Ceará, campus do Pici, 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.
In response to the growing challenges faced by agricultural production, increasing productivity in already cultivated areas has become essential to ensure global food security. In this context, the use of high-vigor seeds is crucial to achieving higher crop yields. However, traditional vigor assessment methods are time-consuming, often manual, and dependent on specialized labor, which drives the search for automated solutions based on Computer Vision and Deep Learning techniques. Several studies propose systems that extract morphological metrics from seedlings, such as the length of their structures and other physiological characteristics, with accurate segmentation of these structures being critical for the reliability and repeatability of the results. This study aimed to evaluate and compare different architectures derived from U-Net for the semantic segmentation of soybean seedlings, considering the three main morphological structures of the seedling: cotyledon, hypocotyl, and primary root. For comparative evaluation purposes, the models analyzed included U-Net, V-Net, RAUNet, UNeXt, WRANet, Swin-Unet, and TransUNet. The dataset consisted of 80 images of seedlings three days after the germination test, expanded to 160 through data augmentation techniques. For a comprehensive performance analysis, evaluation metrics included Precision, Recall, Dice, and Intersection over Union (IoU), as well as complexity indicators such as number of parameters, GFLOPs, and inference time, providing a comprehensive view of both performance and computational efficiency. The results indicated that WRANet achieved the best quantitative performance (Precision: 87.4%; Recall: 92.3%; Dice: 89.7%; IoU: 82.2%), particularly excelling in the segmentation of the most complex class identified in this study, the hypocotyl. However, it

presented the highest inference time, limiting its applicability in real-time processing scenarios. Otherwise, U-Net and TransUNet demonstrated competitive performance, with metrics close to those of WRANet and lower computational cost, making them more balanced and feasible alternatives for practical applications in seed vigor assessment and automated soybean seedling monitoring.

Keywords.
Semantic segmentation; Soybean seedling; Seed vigor; Deep Learning.

Introduction

The agricultural production systems face increasingly complex challenges, driven by longstanding pressures such as rapid population growth, intensified urbanization, and increasing competition for natural resources (Godfray et al., 2010). These factors are further exacerbated by the impacts of climate change, which intensify the instability of production processes and increase the vulnerability of the most exposed populations (Wheeler and von Braun, 2013).

In addition, the availability of land suitable for agriculture is limited, particularly when considering the need to avoid large-scale deforestation and the occupation of marginal areas (Smith et al., 2010).

In this context, expanding the agricultural frontier has become an increasingly less viable and environmentally unsustainable strategy. Thus, evidence suggests that the most effective way to ensure an adequate food supply for a growing population is to increase the productivity of already cultivated areas and improve crop management, rather than expanding land allocated for production (Fróna et al., 2019).

One of the most effective strategies for increasing agricultural productivity is the use of high-vigor seeds (Medeiros et al., 2020; Zhang et al., 2023; Zou et al., 2023). Seed quality and vigor, as a crucial link in the production chain, directly influence plant emergence, early establishment, growth, and consequently, crop yield and quality (Yang et al., 2026; Ali et al., 2021).

Thus, seed companies have been increasingly in demand, promoting the greater use of certified seeds with declared quality (Win et al., 2022). Currently, the seed industry primarily evaluates the physiological quality of seed lots through vigor tests, since high levels of vigor are directly associated with the potential to enhance growth and productivity in agricultural production (Han et al., 2014).

However, traditional vigor assessment methods may require one or more weeks to be completed, in addition to involving labor-intensive procedures dependent on specialized personnel (Wen et al., 2018). Given these limitations, the adoption of automated methods becomes necessary. In this context, Computer Vision stands out by enabling the extraction and interpretation of information from images in a fast, objective, and scalable manner.

Several studies have proposed automated systems based on Digital Image Processing for seed vigor assessment. Many of these systems estimate physiological potential from metrics extracted from seedling images, such as seedling structure length, enabling the objective quantification of vigor (Castan et al., 2018; Hoffmaster et al., 2005; Sako et al., 2001).

However, measuring these lengths requires proper segmentation of seedling structures, a step recognized as one of the most critical in Computer Vision systems, as its quality directly influences the performance and reliability of the obtained results (Xu, 2024).

Therefore, precise and consistent segmentation constitutes an essential condition for the accurate extraction of morphological features and, consequently, for the proper classification of seedling vigor. Considering the complexity of the task, which involves distinguishing among three morphological classes (cotyledons, hypocotyl, and root), this study aims to evaluate and compare

the performance of different architectures from the U-Net family (U-Net with batch normalization, V-Net, RAUNet, UNeXt, WRANet, Swin-Unet, and TransUNet), in order to identify the approach with the greatest suitability for the automatic segmentation of soybean seedlings.

Family of U-Net architectures

The U-Net is known for its “U”-shaped structure, in which the downsampling path is referred to as the encoder and the upsampling path as the decoder. In the encoder, feature maps are generated, increasing channel depth while reducing image dimensions. In the decoder, the image is reconstructed to its original size by continuously reducing the layers generated during the encoding process. Furthermore, the main innovation introduced by U-Net is the use of skip connections, which integrate features obtained from the encoder path while reconstructing the image in the decoder, providing more detailed localization information (Ronneberger et al., 2015).

These distinctive characteristics of the 2015 U-Net constitute the foundation of derived architectures in the literature. The family of models based on U-Net incorporates structural modifications and new components into the classical architecture, aiming to achieve significant improvements in performance and generalization across multiple contexts (Wang et al., 2025; Gao et al., 2025). In summary, the architectures used in the present study, as well as their main additions and particular characteristics, are presented in Table 1.

Table 1. Summary of the models used in the present study.

Architectures	Major additions or changes	Building block paradigm
U-Net com Batch Normalization	Batch normalization	CNN
V-Net	Replacement of ReLU with PReLU activation + 5×5 convolutions	CNN
RAUNet	ResNet34 backbone with attention mechanisms in skip connections	CNN
UNeXt	Feature tokenization for an MLP in deeper layers	CNN + Techniques shared with the Transformer
WRANet	Reduced use of convolutions + wavelet transform replacing max pooling + residual attention module	CNN
Swin-Unet	Replacement of convolutions with Transformer blocks + specialized attention modules	Transformer
TransUnet	Transformer block at the U-Net bottleneck	CNN + Transformer

Derived architectures

Zhou and Yang (2018) proposed a modified U-Net with batch normalization (BN) layers in the convolutional blocks of the classical U-Net architecture proposed by Ronneberger et al. (2015) and found that this single addition adapted the architecture to achieve more stable training and superior results. The U-Net architecture with the addition of BN is shown in Figure 1. Given the above, this modification proposed by Zhou and Yang was chosen to ensure a fair comparison,

since more modern CNNs adopt batch normalization as a standard component of their architectures.

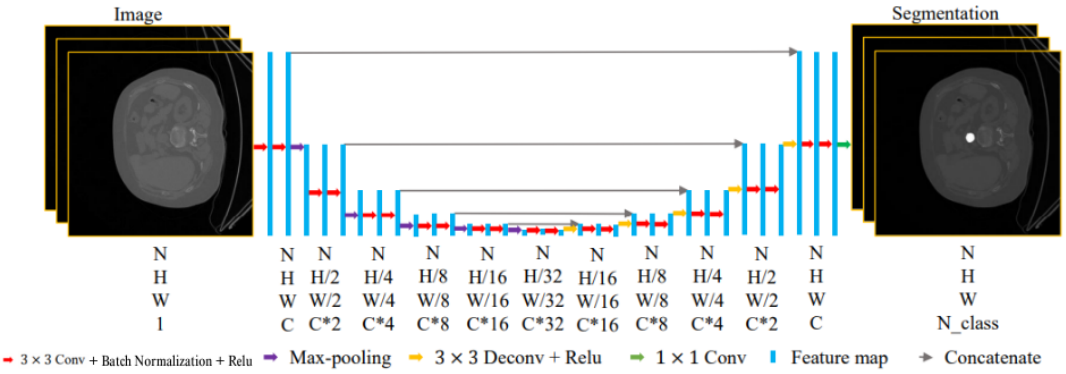

Fig 1. U-Net architecture with Batch Normalization. Adapted from Zhou e Yang (2018).

Milletarì et al. (2016) proposed V-Net, an architecture that replaces the original activation function of U-Net (ReLU) with PReLU. The convolution kernel size was increased from 3x3 to 5x5, a 2x2 convolution was added after each downsampling operation, and residual connections were incorporated into each convolutional block, ensuring faster model convergence during training. The final arrangement of the architecture's blocks is shown in Figure 2d.

Ni et al. (2019) developed the RAUNet architecture by replacing the original U-Net backbone with the ResNet34 backbone (He et al., 2016), pre-trained on the ImageNet dataset. In addition, an augmented attention module was employed to emphasize regions of interest (RoIs) and improve feature representation after the skip connections. These additions are illustrated in Figure 2a.

The UNeXt architecture, proposed by Valanarasu and Patel (2022), was designed as a lightweight alternative, reducing the number of convolutions in the first three layers to a single 3×3 convolution. For the last two layers, the authors proposed replacing each block with a single module named tokenized multi-layer perceptron (Tok-MLP). The region where these Tok-MLP blocks were placed can be observed in Figure 2c. Similar to architectures containing Transformer blocks, this module employs a feature tokenization technique applied to the features extracted from the previous layer, which consists of shifting the image along the width and height dimensions and compressing it in order to apply the MLP process on 3×3 kernels of this compressed representation. Through these processes, the architecture is able to efficiently represent feature maps while requiring minimal computational resources and maintaining low inference time.

In Zhao et al. (2023), the WRANet architecture was developed, with its main contribution consisting of the use of wavelet transform techniques as structural blocks, replacing the max pooling layers in the encoder region of the architecture. This technique employs high-pass and low-pass filters, enabling the layer to extract features more efficiently while simultaneously removing unnecessary image noise arising from high-frequency components by decomposing an image into low- and high-frequency components.

In addition, the authors proposed the inclusion of an attention module in the skip connections using a residual technique, referred to as the Residual Attention Module, in order to improve the decoder stage by emphasizing regions with more relevant features. The contributions proposed by Zhao et al. (2023) can be observed in Figure 2b.

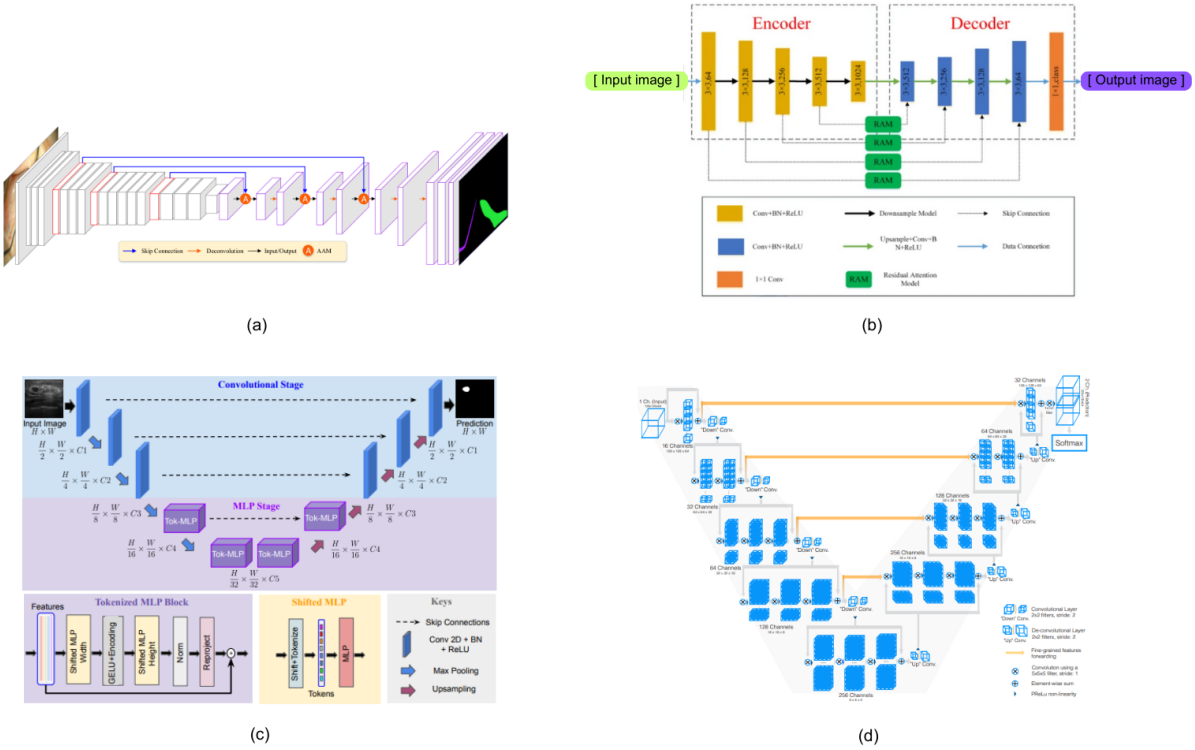

Fig 2. Architectures that do not employ Transformer blocks. a) RAUNet architecture proposed by Ni et al. (2019), b) WRANet architecture adapted from Zhao et al. (2023), c) UNeXt architecture proposed by Valanarasu and Patel (2022), d) V-Net architecture proposed by Milletari et al. (2016).

The main component of the Swin-Unet (Cao et al., 2021) (Figure 3a) and TransUNet (Chen et al., 2021) (Figure 3b) architectures lies in the use of Transformer layers, which were originally designed for natural language processing but later adapted to the domain of CNNs (Vaswani et al., 2017; Turner, 2023).

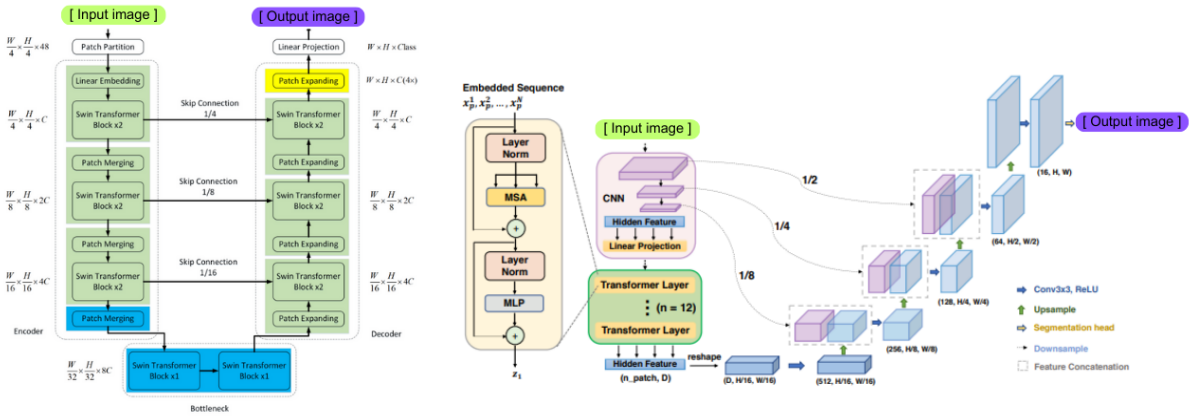

Fig 3. Architectures employing Transformer blocks. a) Swin-Unet architecture adapted from Cao et al. (2021), b) TransUNet architecture adapted from Chen et al. (2021).

The Swin-Unet (Cao et al., 2021) is the first architecture in the U-Net family to be entirely composed of Transformer-based blocks, meaning that the encoder, bottleneck, and decoder all incorporate this method for feature extraction. These components are built using Swin Transformer blocks. Such blocks consist of a LayerNorm (LN) layer, a multi-head self-attention module, a residual connection, and a two-layer MLP with a GeLU activation function. Specifically, the window-based multi-head self-attention (W-MSA) module and the shifted window-based multi-

head self-attention (SW-MSA), both attention modules employed in the architecture, were developed to incorporate attention mechanisms into the Transformer framework while reducing computational cost and preserving model performance. In summary, an input tensor is divided into smaller parts (patches), which undergo a prior tokenization stage before feeding the Swin blocks, enabling the architecture to achieve deep feature representations at different levels. In this way, this convolution-free approach can better learn both global and deep interactions, resulting in improved segmentation of the target object.

In contrast, TransUNet (Chen et al., 2021) is the first framework developed for medical image segmentation that employs a hybrid CNN–Transformer architecture. To compensate for the loss of high-resolution features when using Transformers, the model leverages both the high-resolution spatial information extracted by CNNs and the global context encoded by Transformers through self-attention mechanisms. The self-attention mechanisms are introduced into the encoder path of the U-Net immediately after the feature map generation process is completed. Subsequently, the feature patches are adjusted as input to the Transformer process, which consists of multi-head self-attention layers followed by a multi-layer perceptron (MLP) block. Finally, the architecture proceeds with the standard U-Net decoder process until the final image reconstruction stage.

Materials and Methods

Dataset

The dataset used in the present study consisted of a set of 80 RGB images with a resolution of 4000 × 3000 pixels of soybean seedlings belonging to the Leguminosae family, Papilionoideae subfamily, and *Glycine* L. genus. These images contained 50 seedlings at different developmental stages, collected three days after the germination test had been conducted.

The Roboflow platform, developed for collaborative dataset management and creation, was used to annotate the images considering the three fundamental morphological structures of a seedling: cotyledon, hypocotyl, and root. Each region of interest (RoI) was labeled using a polygon selection tool so that the dataset would be suitable for the semantic segmentation task. After annotation, the images were resized to 640 x 640 pixels to improve the training efficiency of the U-Net models and also underwent an auto-orient process aimed at reducing errors between annotation locations and their corresponding RoIs in the RGB images.

Regarding the dataset, data augmentation techniques were also applied to the images. This synthetic augmentation introduces variability and increases dataset volume, significantly improving the overall generalization capability of the models. Thus, the techniques employed in the present study included horizontal and vertical flips of 180°, 90° rotations (clockwise or counterclockwise), random rotations of ±10°, and the addition of 0.25% noise, resulting in a total of 160 soybean seedling images.

Experiment Design

The present study made use of the Google Colab platform, a Google service designed for hosting and executing scripts in a cloud computing environment. The virtual machine (VM) configured for the experimental development features an Intel® Xeon® CPU @ 2.20 GHz processor with 12 logical cores, x86_64 architecture, 52 GB of RAM, and an NVIDIA T4 GPU with 15 GB of VRAM and support for CUDA 12.4.

The algorithms were developed in the Colab environment using a Jupyter Notebook, enabling the execution of Python code. The architectures were implemented using the PyTorch library, a framework specifically designed and optimized for the development of Machine Learning algorithms.

Regarding the training stage, the following initialization parameters were defined: 150 epochs, the AdamW optimizer, and the learning rate $1x10^{-4}$ and a batch size of 4. It is worth noting that WRANet was the only model that required training with a reduced batch size. This occurred due to the high complexity of the model, which demanded more RAM than was available when values greater than 1 were selected. However, the remaining parameter values were kept consistent across all evaluated models.

Additionally, the dataset was split into 80% of the images for training, 10% for validation, and 10% for testing. The loss function adopted in this study consisted of a weighted sum of the cross-entropy (CE) loss and the Dice loss, whose formulation is presented in Equation 1.

$$Loss_{combo} = (0.5 \times Loss_{CE}) + (0.5 \times Loss_{Dice}) \quad (1)$$

Thus, the $Loss_{combo}$ prevents the models from ignoring underrepresented classes, allowing the model weights to be updated in a way that better reflects the complex nature of the problem, thereby improving the overall performance of the trained models.

Evaluation Metrics

Regarding performance evaluation, the metrics were selected to best describe the quality of the regions segmented by the models. Precision (Equation 2) represents the proportion of pixels predicted as belonging to a class that actually belong to it; recall (Equation 3) refers to the pixels correctly identified for a given class; the Dice coefficient (Equation 4) measures the degree of similarity between two sets of pixels, while the Intersection over Union (IoU) metric (Equation 5) quantifies the overlap between two pixel regions.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Where (TP) represents the number of true positives, (TN) denotes the number of true negatives, (FP) indicates the number of false positives, and (FN) refers to the number of false negatives.

$$Dice = \frac{2|A \cap B|}{|A|+|B|} \quad (4)$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

Where A refers to a region and B corresponds to the same region predicted by the model for a given class.

The number of Giga Floating Point Operations Per Second (GFLOPs), millions of parameters, and inference time were defined as auxiliary metrics, guided by model efficiency according to measures of complexity and execution time. GFLOPs represent the number of arithmetic operations required for the model to process an input. The number of parameters corresponds to the total number of trainable weights in the neural network. Models with more parameters generally possess greater representation capacity, but also a higher risk of overfitting, that is, memorizing the training set without properly generalizing. Inference time corresponds to the interval required for the model to process an input and generate its output, reflecting the practical execution speed of the algorithm.

Results and Discussion

The average metric results during the model validation stage are presented in Table 2. It can be observed that the only model to perform substantially below expectations was Swin-Unet, which may have occurred due to its paradigm shift, as it does not rely on convolutions for feature extraction. WRANet stood out as the best-performing model in this analysis, outperforming all other models despite presenting a reduced number of convolution operations. This suggests that the use of wavelet transformation and residual attention may have been fundamental in achieving values of 0.874, 0.923, 0.897, and 0.822 for Precision, Recall, Dice, and mIoU, respectively, within the domain of soybean seedling images.

Table 2. Qualitative results of the U-Net-derived models during the validation stage.

Architectures	Metrics			
	Precision	Recall	Dice	mIoU
U-Net + BN	0.874	0.905	0.888	0.817
V-Net	0.592	0.908	0.681	0.556
RAUNet	0.819	0.866	0.836	0.746
UNeXt	0.584	0.704	0.591	0.490
WRANet	**0.874**	**0.923**	**0.897**	**0.822**
Swin-Unet	0.486	0.592	0.504	0.413
TransUnet	0.858	0.899	0.877	0.800

Similarly, regarding the performance of the mean IoU metric per class, an analysis of Table 3 reveals that WRANet achieved superior values even for the hypocotyl class, which is perceptibly the most complex within this context, since other models presented unsatisfactory results for this same class. The exceptions were U-Net with BN and TransUNet, which showed results close to those of WRANet across all classes. Therefore, considering only the performance metrics, the three best-performing models can be considered to be WRANet, U-Net with BN, and TransUNet.

Table 3. Mean IoU results per class for each architecture during the validation stage.

Architectures	Class		
	Cotyledon	Root	Hypocotyl
U-Net + BN	0.870	0.753	0.656
V-Net	0.651	0.392	0.277
RAUNet	0.799	0.649	0.558
UNeXt	0.624	0.282	0.209
WRANet	**0.883**	**0.757**	**0.659**
Swin-Unet	0.447	0.204	0.115
TransUnet	0.855	0.734	0.624

It can be observed that, among the attributes compared in Table 4, none of the models emerges as a specialist across all complexity-related aspects. As previously discussed, WRANet qualifies as the best model in terms of performance metrics evaluation; however, it exhibits high complexity, since its inference time is longer than that of the other models and, additionally, it possesses the highest number of GFLOPs.

Table 4. Comparison of the complexity of the models used in the study.

Architectures	Parameters(M) ↓	GFLOPs ↓	FPS ↑	Inference time (ms) ↓
U-Net + BN	31.038	342.190	4.47	223.694
V-Net	9.358	92.339	4.78	209.021
RAUNet	22.143	43.996	**21.96**	**45.532**
UNeXt	14.306	**43.060**	9.72	102.899
WRANet	**3.227**	561.086	2.15	465.594
Swin-Unet	27.146	48.329	15.15	65.140
TransUnet	312.90	453.138	3.35	298.950

Considering only performance values, the WRANet architecture would be selected as the best due to its higher efficiency. However, when real-time applications are taken into account, this architecture becomes a less favorable choice, since its inference time of 465.594 ms falls short of what is considered satisfactory for real-time requirements. In scenarios where latency is a

critical factor, U-Net with BN and TransUNet emerge as better options, as their metric values are numerically close to those achieved by WRANet.

Conclusion

The present study aimed to evaluate a set of deep learning models derived from the U-Net architecture available in the literature, which were trained using soybean seedling images collected three days after the germination test. The results obtained in this study show that, among the evaluated models, none was superior across all assessment measures, meaning that the best-performing models in terms of quantitative performance metrics did not necessarily correspond to the best results in terms of computational complexity.

In general terms, the U-Net architecture and its variants achieved average values of 72.67% (± 15.38%), 82.81% (± 11.88%), 75.34% (± 14.86%), and 66.34% (± 15.97%) for Precision, Recall, Dice, and IoU, respectively, indicating the practical potential of their use in RGB images for the semantic segmentation of soybean seedling structures.

The model derived from the WRANet architecture stood out as the one that numerically achieved the most satisfactory results during the evaluation stage, with values of 87.4% for Precision, 92.3% for Recall, 89.7% for Dice, and 82.2% for IoU. However, in terms of computational demand, WRANet was identified as the slowest model (highest inference time) compared to the others.

It is worth noting that the TransUNet model achieved 85.8% Precision, 89.9% Recall, 87.7% Dice, and 80.0% IoU, while U-Net with BN achieved 87.4% Precision, 90.5% Recall, 88.8% Dice, and 81.7% IoU. Both models stood out positively by presenting competitive results when compared to WRANet, including in terms of IoU for the hypocotyl class, identified as the most complex class throughout the study. From a computational efficiency perspective, although both models present inference times above 200 ms per image, TransUNet and U-Net with BN exhibit performance metrics relatively close to WRANet, making them more suitable alternatives for real-time applications.

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

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., & Toulmin, C. (2010). Food Security: the Challenge of Feeding 9 Billion People. *Science*, 327(5967), 812–818. https://doi.org/10.1126/science.1185383

Wheeler, T., & von Braun, J. (2013). Climate Change Impacts on Global Food Security. *Science*, 341(6145), 508–513. https://doi.org/10.1126/science.1239402

Smith, P., Gregory, P. J., van Vuuren, D., Obersteiner, M., Havlík, P., Rounsevell, M., Woods, J., Stehfest, E., & Bellarby, J. (2010). Competition for land. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1554), 2941–2957. https://doi.org/10.1098/rstb.2010.0127

Fróna, D., Szenderák, J., & Harangi-Rákos, M. (2019). The Challenge of Feeding the World. *Sustainability*, 11(20), 5816.

Medeiros, A. D. de, Silva, L. J. da, Ribeiro, J. P. O., Ferreira, K. C., Rosas, J. T. F., Santos, A.

A., & Silva, C. B. da. (2020). Machine Learning for Seed Quality Classification: An Advanced Approach Using Merger Data from FT-NIR Spectroscopy and X-ray Imaging. *Sensors*, 20(15), 4319. https://doi.org/10.3390/s20154319

Zhang, T., Lu, L., Yang, N., Fisk, I. D., Wei, W., Wang, L., Li, J., Sun, Q., & Zeng, R. (2023). Integration of hyperspectral imaging, non-targeted metabolomics and machine learning for vigour prediction of naturally and accelerated aged sweetcorn seeds. *Food Control*, 153, 109930–109930. https://doi.org/10.1016/j.foodcont.2023.109930

Zou, Z., Chen, J., Wu, W., Luo, J., Long, T., Wu, Q., Wang, Q., Zhen, J., Zhao, Y., Wang, Y., Chen, Y., Zhou, M., & Xu, L. (2023). Detection of peanut seed vigor based on hyperspectral imaging and chemometrics. *Frontiers in Plant Science*, 14. https://doi.org/10.3389/fpls.2023.1127108

Yang, T., Peng, B., Zhang, J., Zhang, D., Liang, C., & Fan, X. (2026). Optimizing seed anomaly detection in agricultural automation via lightweight ASD-YOLO and closed-loop control. *Computers and Electronics in Agriculture*, 243, 111342. https://doi.org/10.1016/j.compag.2025.111342

Ali, A., Qadri, S., Mashwani, W. K., Brahim Belhaouari, S., Naeem, S., Rafique, S., Jamal, F., Chesneau, C., & Anam, S. (2020). Machine learning approach for the classification of corn seed using hybrid features. *International Journal of Food Properties*, 23(1), 1110–1124. https://doi.org/10.1080/10942912.2020.1778724

Win, M. T., Maredia, M. K., & Boughton, D. (2022). Farmer demand for certified legume seeds and the viability of farmer seed enterprises: Evidence from Myanmar. *Food Security*. https://doi.org/10.1007/s12571-022-01338-0

Han, Z., Ku, L., Zhang, Z., Zhang, J., Shu Lei Guo, Liu, H., Zhao, R., Ren, Z., Zhang, L., Su, H., Dong, L., & Chen, Y. (2014). QTLs for Seed Vigor-Related Traits Identified in Maize Seeds Germinated under Artificial Aging Conditions. *PLOS ONE*, 9(3), e92535–e92535. https://doi.org/10.1371/journal.pone.0092535

Wen, D., Hou, H., Meng, A., Meng, J., Xie, L., & Zhang, C. (2018). Rapid evaluation of seed vigor by the absolute content of protein in seed within the same crop. *Scientific Reports*, 8(1). https://doi.org/10.1038/s41598-018-23909-y

Castan, D. O. C., Gomes-Junior, F. G., & Marcos-Filho, J. (2018). Vigor-S, a new system for evaluating the physiological potential of maize seeds. *Scientia Agricola*, 75(2), 167–172. https://doi.org/10.1590/1678-992x-2016-0401

Hoffmaster, A. F., Xu, L., Fujimura, K., Bennett, M. A., Evans, A. F., & McDonald, M. B. (2005). The Ohio State University Seed Vigor Imaging System (SVIS) for Soybean and Corn Seedlings. *Seed Technology*, 27(1), 7–24. JSTOR. https://doi.org/10.2307/23433211

Sako, Y., McDonald, M. B., Fujimura, K., Evans, A. F., & Bennett, M. A. (2001). A system for automated seed vigour assessment. *Seed Science and Technology*, 29(3), 625–636.

Xu, Y. (2024). Application of Image Segmentation Algorithms in Computer Vision. *Frontiers in Computing and Intelligent Systems*, 7(3), 17–20. https://doi.org/10.54097/gq1s6737

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241), Springer: Cham, Germany

Wang, J., Ruhaiyem, N.I., & Fu, P. (2025). A Comprehensive Review of U-Net and Its Variants: Advances and Applications in Medical Image Segmentation. *ArXiv*, abs/2502.06895.

Gao, Y., Jiang, Y., Peng, Y., Yuan, F., Zhang, X., & Wang, J. (2025). Medical Image

Segmentation: A Comprehensive Review of Deep Learning-Based Methods. *Tomography*, 11(5), 52. https://doi.org/10.3390/tomography11050052

Zhou, X., & Yang, G. (2018). Normalization in Training U-Net for 2-D Biomedical Semantic Segmentation. *IEEE Robotics and Automation Letters*, 4, 1792-1799.

Milletarì, F., Navab, N., & Ahmadi, S. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision* (3DV), 565-571.

Ni, Z., Bian, G., Zhou, X., Hou, Z., Xie, X., Wang, C., Zhou, Y., Li, R., & Li, Z. (2019). RAUNet: Residual Attention U-Net for Semantic Segmentation of Cataract Surgical Instruments. *ArXiv*, abs/1909.10360.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Valanarasu, J., & Patel, V.M. (2022). UNeXt: MLP-based Rapid Medical Image Segmentation Network. *ArXiv*, abs/2203.04967.

Zhao, Y., Wang, S., Zhang, Y., Qiao, S., & Zhang, M. (2023). WRANet: wavelet integrated residual attention U-Net network for medical image segmentation. *Complex & Intelligent Systems*, 9(6), 6971–6983. https://doi.org/10.1007/s40747-023-01119-y

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *ECCV Workshops*. https://doi.org/10.48550/arxiv.2105.05537

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., & Zhou, Y. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *ArXiv*, abs/2102.04306.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Neural Information Processing Systems*.

Turner, R.E. (2023). An Introduction to Transformers. *ArXiv*, abs/2304.120557.