Graph-Based Semantic Segmentation of Agricultural Images Using CNN Features, Vegetation Spectral Indices, and Wavelet Texture Descriptors

Esther Gascó[1][0000-0003-1798-4483]*, Clara I. López-González[2][0000-0002-2755-1692], Eva Besada-Portas[2][0000-0001-6129-4653] and Gonzalo Pajares[3][0000-0003-0915-6282]

[1] Dpt. of Software Engineering and Artificial Intelligence, Complutense University of Madrid, Madrid, 28040, Spain, egasco@ucm.es
[2] Dpt. of Computer Architecture and Automation, Complutense University of Madrid, Madrid, 28040, Spain, claraisl@ucm.es, ebesada@ucm.es
[3] Institute for Knowledge Technology, Complutense University of Madrid, Madrid, 28040, Spain, pajares@ucm.es
* Corresponding author

Abstract. This study proposes a graph-based methodology for automatic vegetation-soil discrimination in agricultural images. The method combines convolutional visual features with spectral information derived from the Excess Green (ExG) vegetation index, multi-scale texture descriptors obtained from the Discrete Wavelet Transform (DWT), and contextual modelling through Graph Convolutional Networks (GCNs). Visual features are first generated by a ResNet-18 encoder and subsequently fused with the handcrafted descriptors within SLIC superpixels to build a graph representation of the scene. The resulting graph is processed by a two-layer GCN to classify vegetation and soil regions. Experimental evaluation on a dataset of 1,373 labelled agricultural images yields a Mean Accuracy of 0.94 and an mIoU of 0.85, outperforming several established segmentation methods. These results show that combining spectral, textural, and graph-based representations improves vegetation-soil discrimination in realistic agricultural environments.

Keywords: Vegetation detection, Semantic segmentation, Graph convolutional networks, Wavelet texture analysis.

1 Introduction

Accurate semantic segmentation of agricultural imagery is a key requirement for precision farming, enabling applications such as yield estimation, crop inspection, and autonomous vehicle guidance. In row-crop scenarios, reliable separation of vegetation and soil constitutes a fundamental step for field monitoring and decision-making. Nevertheless, images acquired under practical farming conditions are affected by variations in illumination, soil appearance, crop development, and plant occlusions, making robust vegetation-soil discrimination a challenging task.

The rapid evolution of deep learning has considerably advanced semantic image segmentation, with Convolutional Neural Networks (CNNs) becoming the dominant approach. Representative architectures, including Fully Convolutional Networks (FCN) [1], SegNet [2], U-Net [3], and DeepLab [4], have achieved remarkable performance in vision-based agricultural tasks, including crop surveillance, plant disease assessment, and vegetation characterization [5,6]. Despite these advances, conventional CNNs primarily exploit local spatial information, making the modelling of long-range contextual dependencies more difficult.

Graph-based representation learning, implemented here through a Graph Convolutional Network (GCN), address this limitation by modelling image regions as graph nodes connected through spatial relationships, enabling contextual information propagation. Recent studies have shown that combining CNN features with graph reasoning improves plant disease detection and crop monitoring [7-9].

Agricultural imagery also contains distinctive spectral and textural information. Vegetation-enhancing representation is generated through the Excess Green (ExG) index [10], whereas texture is represented using multi-scale Discrete Wavelet Transform (DWT) coefficients [11].

Building upon these observations, the proposed method combines complementary deep, spectral, texture, and graph-based representations into a unified semantic segmentation pipeline for vegetation-soil discrimination. Rather than proposing a new segmentation architecture, the main contribution lies in the effective integration of these complementary representations within a graph-based framework. CNN, ExG, and DWT features are aggregated over SLIC superpixels to construct an image graph, which is subsequently processed by a two-layer GCN to model contextual relationships between neighbouring regions. The main contributions are:

1. A hybrid CNN-ExG-DWT-GCN framework integrating complementary deep, spectral, texture, and graph-based representations for vegetation-soil segmentation.
2. A superpixel-based graph representation aggregating CNN, ExG, and DWT features into compact node descriptors.
3. A two-layer GCN that performs contextual reasoning between neighbouring superpixel regions.
4. An experimental validation demonstrating the computational efficiency of the proposed framework together with its ability to satisfy real-time processing requirements for agricultural robotic systems.

Section 2 describes the proposed methodology, Section 3 details the experimental setup, and Section 4 presents and discusses the obtained results.

2 Overview of the Proposed Architecture

The proposed methodology addresses vegetation-soil discrimination in row-crop imagery by combining complementary sources of visual information. The overall processing pipeline CNN-ExG-DWT-GCN, illustrated in Figure 1, consists of four sequential stages:

1. CNN feature extraction. A ResNet-18 backbone extracts dense semantic feature maps from the input image.
2. Spectral and texture feature extraction. Spectral information is obtained using the Excess Green (ExG) vegetation index, whereas texture is characterized using multi-scale Discrete Wavelet Transform (DWT) coefficients.
3. Superpixel-based graph construction and feature fusion. The graph nodes are obtained from an SLIC over-segmentation of the input image [12]. CNN, ExG, and DWT features are aggregated within each superpixel and concatenated to form the node descriptors, while graph edges encode spatial adjacency.
4. Graph-based node classification. A two-layer GCN propagates contextual information between neighbouring nodes to produce the final vegetation-soil segmentation.

The proposed framework is conceptually related to our previous hybrid architecture [13], as both combine deep, spectral, and region-based representations using the same ResNet-18 backbone and SLIC-based superpixel partition. However, the present work replaces the Sobel structural branch and attention head of [13] with DWT-based texture descriptors and a GCN module, enabling explicit graph-based contextual reasoning between neighbouring regions.

The proposed methodology combines deep visual representations with spectral, textural, and graph-based information to jointly exploit local image characteristics and contextual relationships for vegetation-soil discrimination.

2.1 CNN Backbone for Feature Extractor

Visual features are extracted using an ImageNet pre-trained ResNet-18 encoder. The residual learning strategy enables effective gradient propagation, facilitating the optimization of deep networks while avoiding the degradation problems of conventional CNNs. Unlike earlier CNN architectures, including AlexNet and VGG or ResNet-18 offers a favourable balance between representational capability and computational efficiency, making it well suited for real-time agricultural vision applications. In addition, using a pre-trained backbone allows the model to exploit generic visual representations learned from large-scale image collections, which improves convergence and generalization when training on relatively small agricultural datasets.

ResNet-18 was preferred over deeper variants such as ResNet-50 and ResNet-101 because it provides an effective compromise between feature representation and computational efficiency. Although deeper backbones increase modelling capacity, they also incur higher computational cost with only marginal performance improvements for the present task. In the proposed framework, feature maps are further processed at the superpixel level and propagated through a graph convolutional network, where the computational complexity scales with the number of graph nodes and the dimension of their feature vectors. Moreover, the target application of this work involves deployment in vision systems for autonomous agricultural vehicles, where fast inference under resource-constrained hardware is essential. Therefore, a lightweight backbone such as ResNet-18 provides sufficiently rich semantic features while keeping the overall pipeline efficient and suitable for embedded or on-board processing.

The classification head of the original ResNet-18 model is removed, since the network is used as a dense feature extractor rather than as an image classifier. Given an input RGB image $I \in \mathbb{R}^{512\times512\times3}$, the CNN produces a feature representation: $F \in \mathbb{R}^{H\times W\times C}$ where H and W correspond to the height and width (spatial dimensions) of the feature map and C the feature dimensionality set to 128 as a compromise to retain spatial information while encoding high-level semantic CNN-derived features.

To adapt the ResNet backbone to the proposed graph-based segmentation framework, a 1×1 convolutional projection layer is introduced on the intermediate feature maps generated by the CNN. This layer projects the high-dimensional feature representation into a compact embedding space of dimension C suitable for graph processing. A 1×1 convolution is first applied to project the feature maps into a lower-dimensional representation without altering its spatial structure, improving computational efficiency in the subsequent graph processing stage.

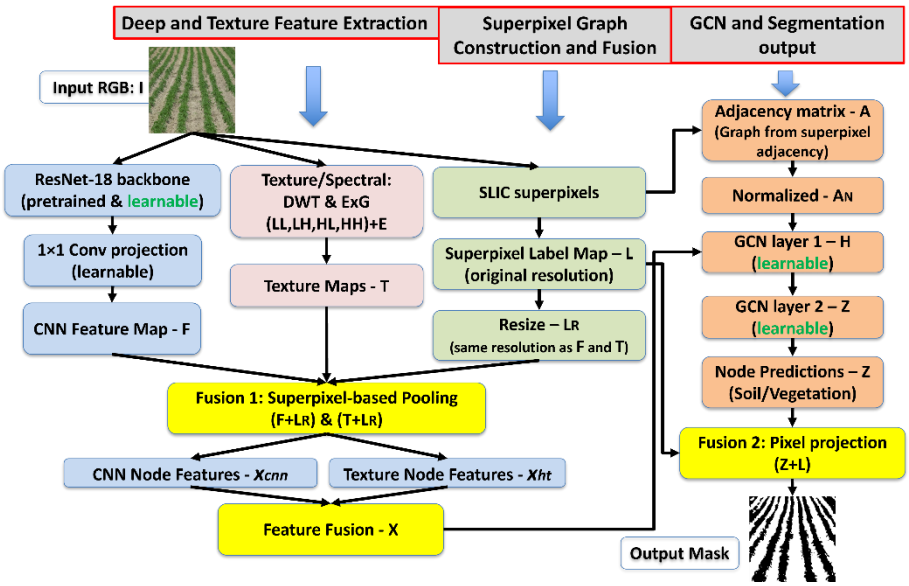


Fig. 1. Proposed DWT-enhanced CNN-GCN architecture. Deep CNN features and wavelet-based texture features are aggregated over SLIC superpixels and fused to form graph nodes, which are processed by a two-layer GCN to obtain the final soil-vegetation segmentation.

2.2 Spectral and Wavelet-based Texture

In addition to the deep visual features extracted by the CNN backbone, the proposed method incorporates handcrafted descriptors aimed at capturing complementary spectral and textural information relevant for vegetation-soil discrimination.

First, spectral information is incorporated through the Excess Green (ExG) vegetation index, defined as ExG = 2G-R-B, where R, G and B denote the red, green, and blue spectral channels of the input image. This transformation enhances the spectral response of vegetation while suppressing the visual contribution of soil and background

regions. The ExG transformation accentuates vegetation-related spectral responses, thereby increasing the contrast with soil and background areas.

Second, a Discrete Wavelet Transform (DWT) is applied to the input image to extract multi-scale texture information. Wavelet analysis provides a compact multi-scale representation of local spatial-frequency patterns, which is particularly useful in agricultural scenes where vegetation and soil exhibit distinct texture characteristics. In this work, a single-level two-dimensional DWT using the Daubechies-4 (db4) wavelet is applied to obtain four sub-band responses corresponding to the approximation (LL) and high-frequency coefficients corresponding to the horizontal (LH), vertical (HL), and diagonal (HH) sub-bands. The approximation coefficients provide a smoothed structural representation of the scene, whereas the detail coefficients capture fine textural variations and local edge information.

The ExG map and the wavelet sub-bands are resized to align the spatial resolution of the CNN feature maps, yielding a handcrafted feature representation $T \in \mathbb{R}^{H \times W \times K}$, where K is the number of wavelet channels. In the proposed implementation $K = 5$, corresponding to one ExG channel and four wavelet-based channels.

By combining ExG with DWT-based descriptors, the proposed representation integrates vegetation-sensitive spectral information with multi-scale texture cues, providing a complementary description to the deep semantic features learned by the CNN.

2.3 Superpixel-based Graph Construction

Rather than operating at the pixel level, the proposed method represents the image as a graph defined over superpixels. Superpixels are generated using the SLIC algorithm, which partitions the image into approximately homogeneous regions while preserving object boundaries. In our experiments, each image is partitioned into approximately 600 SLIC superpixels, yielding a compact representation that preserves the main spatial structures of the scene.

Each superpixel corresponds to a node in the graph. Since the CNN and handcrafted branches produce feature maps at a reduced spatial resolution relative to the original image, the superpixel label map is resized using nearest-neighbour interpolation to align the spatial dimensions of the two feature maps. Nearest-neighbour interpolation is used to preserve the discrete labels of the superpixel regions without introducing interpolation artifacts.

The resized superpixel label map L_R is then used to associate each spatial position of the CNN feature map F and the handcrafted feature map T with its corresponding superpixel region. The image locations assigned to the i-th superpixel are denoted by S_i. Average pooling is applied independently over all positions in S_i to obtain the CNN-based node descriptor and the handcrafted node descriptor as follows:

$$x_i^{cnn} = \frac{1}{|S_i|} \sum_{p \in S_i} F(p) \qquad x_i^{ht} = \frac{1}{|S_i|} \sum_{p \in S_i} T(p) \tag{1}$$

The final node representation is then formed by concatenating both descriptors

$$x_i = \left[x_i^{cnn}, x_i^{ht}\right]; \quad \text{with} \qquad x_i^{ht} = [x_i^{exg}, x_i^{LL}, x_i^{LH}, x_i^{HL}, x_i^{HH}] \tag{2}$$

which yields the node feature matrix $X \in \mathbb{R}^{N\times(C+K)}$ where N denotes the number of superpixels. C denotes the dimensionality of channels in the CNN feature maps, and K is the dimensionality of the handcrafted descriptor. Graph edges are defined according to spatial adjacency between superpixels. Two nodes are connected if their corresponding superpixels share a common boundary in the image. This process results in a graph $G = (V,E)$, where V represents the set of superpixel nodes and E the adjacency relationships between neighbouring regions.

Although ExG may be sensitive to illumination changes and crop variability, its combination with complementary CNN and DWT features improves robustness. Likewise, although the graph representation depends on the initial SLIC partition, the superpixel configuration was selected to preserve object boundaries, providing a compact and reliable graph for contextual reasoning by the GCN.

2.4 Graph Convolutional Network

The concatenated node representations are then fed into a two-layer GCN for contextual reasoning between neighbouring superpixels through the graph structure. This is particularly effective in agricultural scenes, where vegetation and soil exhibit spatially coherent patterns along crop rows.

Let $X \in \mathbb{R}^{N\times(C+K)}$ denote the node feature matrix obtained after feature fusion, where N is the number of superpixels, C is the dimensionality of the CNN features, and K corresponds to the handcrafted spectral-texture descriptors derived from the ExG vegetation index and the wavelet sub-bands. Let A be the adjacency matrix of the graph. The normalized adjacency matrix is computed as

$$A_N = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}} \tag{3}$$

where D corresponds to the degree matrix and I is the identity matrix.

Two graph convolution layers (GCN Layer 1 and 2) are applied to process the graph:

$$H = \sigma(A_N X W_0 + b_0); \quad Z = A_N H W_1 + b_1 \tag{4}$$

where W_0 and W_1 are learnable weight matrices, together with their corresponding bias b_0 and b_1, and $\sigma(\cdot)$ denotes a ReLU activation function. The final output Z produces class logits for each graph node corresponding to the two semantic classes considered in this work: vegetation and soil. These node predictions are then projected back to the pixel domain by assigning the predicted class label of each node to all pixels belonging to the corresponding superpixel region. The hidden dimension of the intermediate graph representation was set to 128, providing sufficient representational capacity while maintaining computational efficiency. Consequently, the graph convolutional layers operate on compact 128-dimensional node embeddings.

A two-layer GCN was adopted to balance contextual modelling and computational efficiency. Although deeper GCNs can capture higher-order relationships, they often suffer from over-smoothing, reducing the discriminative power of node representations. Since superpixel graphs already preserve local spatial structure, two graph convolution layers are sufficient to propagate contextual information between neighbouring regions.

The learnable parameters comprise the weights of the CNN backbone, the 1×1 projection layer, and the graph convolutional head (W_0, b_0, W_1, b_1). In contrast, ExG computation, DWT decomposition, superpixel generation, graph construction, and pixel projection are deterministic and introduce no trainable parameters. The GCN improves segmentation by propagating information between adjacent superpixels, reducing isolated misclassifications near vegetation-soil boundaries and producing spatially coherent regions. This is particularly beneficial in agricultural scenes, where vegetation follows structured crop-row patterns.

3 Experimental setup

Experiments were carried out in MATLAB R2025b (MathWorks) [14] on a workstation featuring an Intel Core i9 CPU, 32 GB of RAM, and an NVIDIA RTX GPU.

The experimental dataset combines the public Crop Row Benchmark Dataset (CRBD) [15] with RGB images collected from agricultural fields in the Vega del Jarama region (Madrid, Spain). In total, 1,373 images acquired under different crop layouts, illumination scenarios, vegetation coverage, and soil appearances were available for evaluation. Integrating both data sources increases scene diversity and improves the robustness of the proposed approach. All images were resampled to 512×512 pixels before being randomly split into training, validation, and test subsets (80/10/10). The partition was generated after merging both datasets, preserving the relative proportion of public and proprietary samples while avoiding overlapping between subsets. The same partition was adopted for every evaluated method.

Reference masks were generated through a human semi-automatic annotation procedure based on K-means clustering as an initial annotation step followed by manual correction.

Model optimization was carried out for 60 epochs with the Adam optimizer, starting from a learning rate of 10^{-4} and applying exponential decay every five epochs with a decay factor of 3×10^{-3}. The optimizer parameters were set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 3\times10^{-7}$. An L2 regularization coefficient of 10^{-3} and a batch size of eight images were employed. The CNN backbone and GCN classifier were optimized jointly in an end-to-end fashion. For graph construction, each image was decomposed into approximately 600 SLIC superpixels using a compactness value of 15.

Segmentation performance was assessed using Mean Accuracy, Mean Intersection over Union (mIoU), Frequency Weighted IoU (FWIoU), and Boundary F-score (BFScore). The mIoU was adopted as the principal evaluation criterion, whereas FWIoU compensates for class imbalance and BFScore measures contour accuracy.

Training employed a weighted graph-node cross-entropy loss. Each node inherited the majority label of its corresponding superpixel, while class weights compensated for the unequal distribution of soil and vegetation samples.

4 Results and Discussions

This section evaluates the proposed framework from complementary perspectives. An ablation analysis is first presented to quantify the contribution of its main components, with particular emphasis on the graph-based reasoning module and the DWT texture representation. The best-performing configuration is then benchmarked against several representative semantic segmentation methods.

4.1 Ablation study

The influence of the main components of the proposed framework was assessed through an ablation study involving several simplified model variants (where S denotes Superpixels): CNN + S, CNN + ExG + DWT + S, CNN + GCN, and the full CNN + ExG + DWT + GCN model. In the non-GCN variants, superpixel descriptors were classified independently using a linear node classifier, whereas in the GCN-based variants contextual reasoning was performed through graph convolution.

The ablation analysis confirms the role of each component in the overall framework. The CNN-based baseline achieves an mIoU of 0.66. Incorporating ExG spectral information together with DWT texture descriptors results in a measurable improvement, raising the mIoU to 0.70 and the BFScore to 0.67, thereby confirming the complementary role of both representations. A further improvement is obtained by introducing the GCN, which exploits contextual relationships between neighbouring superpixels. The CNN+GCN configuration reaches an mIoU of 0.80, demonstrating the effectiveness of graph-based modelling. Finally, the complete architecture integrating CNN features, vegetation spectral indices, wavelet-based texture descriptors, and graph convolution produces the best results, with an mIoU of 0.85 and a BFScore of 0.79, indicating that the integration of spectral, texture, and graph-based contextual information significantly improves boundary localization (Table 1). These results indicate that combining convolutional image features together with spectral, textural, and spatial contextual information significantly improves vegetation-soil segmentation in agricultural scenes.

Table 1. Performance of the different configurations evaluated in the ablation study.

Model variant	Mean Accuracy	mIoU	FWIoU	BFScore
CNN+S	0.83	0.66	0.75	0.61
CNN+ExG+DWT+S	0.84	0.70	0.82	0.67
CNN+GCN	0.89	0.80	0.87	0.73
CNN+ExG+DWT+GCN	**0.94**	**0.85**	**0.93**	**0.79**

The experiments also indicate that ImageNet pre-training increases the mIoU by more than 7% compared with training the network from random initialization. In contrast, replacing ResNet-18 with deeper backbones (ResNet-50 or ResNet-101) yields only marginal gains, indicating that the structural characteristics required for vegetation-soil discrimination are effectively captured by the lightweight ResNet-18 backbone.

The normalized confusion matrix (Table 2) indicates high classification reliability for both categories, with 92.9% of soil pixels and 94.0% of vegetation pixels correctly assigned. The corresponding error rates remain low (7.1% for soil and 6% for vegetation), demonstrating that the combination of spectral, textural, and graph-based representations provides effective separation between vegetation and soil.

From a computational perspective, the proposed CNN-ExG-DWT-GCN framework maintains a moderate complexity despite integrating spectral, texture, and graph-based reasoning modules. As summarized in Table 3, the complete model contains 11.215 million learnable parameters, of which only 17.4 thousand parameters belong to the GCN classification head. This indicates that the vast majority of the model complexity belongs to the pre-trained CNN backbone, whereas the graph reasoning stage introduces only a negligible increase in the number of trainable parameters.

Table 2. Class-normalized confusion matrix for vegetation-soil segmentation.

Class	Soil	Vegetation
Soil	0.929	0.071
Vegetation	0.060	0.940

The approximate computational cost of the graph reasoning module is 208.7 MFLOPs, whereas the additional handcrafted ExG and DWT feature extraction requires only 11.3 MFLOPs, representing a comparatively small overhead with respect to the overall processing pipeline. On the target hardware platform, the proposed framework achieves an average inference latency of 329.2 ± 109.9 ms per 512×512 image, corresponding to approximately 3.04 frames per second (FPS). These results indicate that the proposed framework combines high segmentation performance with moderate computational requirements, making it suitable for embedded agricultural robotic platforms. At a typical operating speed of 5 km/h (≈1.4 m/s), the processing rate of 3.04 FPS provides sufficient throughput for real-time vegetation-soil segmentation.

Table 3. Computational complexity for CNN+ExG+DWT+GCN (512×512 input).

Params (M)	GCN Head (K)	GCN FLOPs (M)	ExG+DWT FLOPs (M)	Latency (ms)	FPS
11.215	17.41	208.7	11.3	329.2 ± 109.9	3.04

Interestingly, the GCN head represents less than 0.2% of the total learnable parameters, confirming that the proposed graph reasoning module introduces only a minimal parameter overhead while providing a measurable improvement in segmentation accuracy.

4.2 Comparison with segmentation architectures

Following the ablation analysis, the highest-performing CNN+ExG+DWT+GCN configuration was evaluated against representative semantic segmentation approaches on the same dataset. Baseline models include FCNNet (following an FCN-8s VGG-

style configuration), SegNet (depth 8), U-Net (with four encoder-decoder levels and skip connections), DeepLabv3+ (ResNet-18), and the CBAM+DWT architecture which integrates a Convolutional Block Attention Module (CBAM) with DWT for plant disease detection, with the aim of capturing complex spatial-spectral relationships [11]. In addition, the recently proposed CNN+ExG+Sobel+Attention framework [13] was included as a closest related approach, since both methods share the same ResNet-18 backbone, ExG spectral branch, and SLIC-based superpixel mapping, differing mainly in the use of DWT instead of Sobel descriptors and a GCN head instead of the attention-based classification module.

For consistency, all methods were assessed using the same dataset partition, input resolution, training protocol, and evaluation metrics. The comparative results are reported in Table 4.

Classical encoder-decoder architectures such as FCNNet and SegNet achieve substantially lower performance than more recent methods, with SegNet obtaining an mIoU of only 0.36 despite being trained under the same optimization protocol. In contrast, DeepLabv3+ and U-Net achieve substantially better results owing to their ability to combine multi-scale representations with rich semantic information. The performance of the CBAM+DWT model is comparable to that of the ablated CNN+ExG+DWT+S configuration, further supporting the contribution of DWT-based texture descriptors.

The comparison with CNN+ExG+Sobel+Attention [13] further isolates the contribution of the proposed modifications. Replacing Sobel descriptors with DWT features produces a modest improvement, whereas replacing the attention head with the GCN leads to the highest overall performance, demonstrating the effectiveness of graph-based contextual reasoning for modelling relationships between neighbouring superpixels.

Table 4. Quantitative benchmarking of the proposed CNN-ExG-DWT-GCN framework against representative segmentation methods.

Model	Mean Accuracy	mIoU	FWIoU	BFScore
FCNNet	0.59	0.48	0.63	0.51
SegNet	0.61	0.36	0.43	0.44
DeepLabv3+ (ResNet18)	0.80	0.78	0.81	0.82
U-Net	0.90	0.76	0.82	0.85
CBAM+DWT	0.92	0.82	0.84	0.81
Hybrid CNN+ExG+Sobel+Att. [13]	0.92	0.83	0.90	0.86
CNN+ExG+DWT+GCN	**0.94**	**0.85**	**0.93**	**0.79**

Overall, the proposed CNN+ExG+DWT+GCN framework achieves the best Mean Accuracy (0.94), mIoU (0.85), and FWIoU (0.93). Although its BFScore (0.79) is lower than those of U-Net (0.85), DeepLabv3+ (0.82), and the hybrid CNN+ExG+Sobel+Attention framework [13] (0.86), this reflects the different optimization characteristics of the compared approaches. The proposed graph-based framework favours regional consistency and contextual classification by propagating infor-

mation between neighbouring superpixels, resulting in superior region-based segmentation performance. In contrast, encoder-decoder architectures and the attention-based head of [13] are more effective at preserving fine boundary details, leading to higher BFScore values. Overall, these results demonstrate that replacing the attention-based head with graph-based contextual reasoning, together with the use of DWT texture descriptors, provides a more robust representation for vegetation-soil segmentation while maintaining competitive boundary localization performance.

4.3 Qualitative analysis

Figure 2 presents the probability maps and the corresponding segmentation generated by the proposed CNN-ExG-DWT-GCN framework. Vegetation is consistently assigned high confidence values following the row structure, while soil areas exhibit low responses. The visual results demonstrate accurate classification together with strong spatial consistency across neighbouring regions.

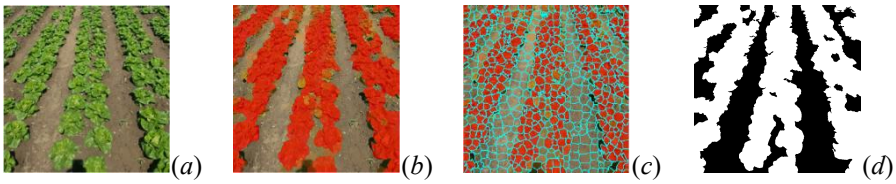


Fig. 2. Representative outputs of the proposed graph-based segmentation framework: (*a*) input RGB image, (*b*) vegetation probability distribution, (*c*) confidence values assigned to superpixel regions, and (*d*) predicted vegetation-soil mask.

5 Conclusions

This work introduced a CNN-ExG-DWT-GCN methodology for vegetation-soil classification in agricultural images by jointly exploiting convolutional image representations, ExG spectral information, DWT-based texture analysis, and graph-based contextual modelling. Feature descriptors are aggregated at the superpixel level to construct a graph, where a two-layer GCN captures spatial interactions between neighbouring regions and improves segmentation consistency.

Experimental results demonstrate that the proposed methodology consistently outperforms representative semantic segmentation models, confirming that the combination of spectral, textural, and graph-based information provides an effective representation for vegetation-soil discrimination under challenging field conditions.

Future research will explore additional vegetation indices together with more advanced graph learning techniques to further enhance robustness across different crop species and acquisition scenarios.

Acknowledgments. This work has been supported by the Research Project BEST-WATER under the Grant PID2024-156033OB-C31 funded by MICIU/AEI/ 10.13039/501100011033 and by "ERDF/EU". The authors thank the anonymous referees for their very valuable comments and suggestions.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431-3440 (2015)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12), 2481-2495 (2017)
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234-241. Springer, Heidelberg (2015)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833-851. Springer, Heidelberg (2018)
5. Sodjinou, S. G., Mohammadi, V., Mahama, A.T.S., Gouton, P.: A deep semantic segmentation-based algorithm for crop and weed detection. Computers and Electronics in Agriculture 196, 106873 (2022)
6. Luo, Z., Yang, W., Yuan, Y., Gou, R., Li, X.: Semantic segmentation of agricultural images: A survey. Information Processing in Agriculture 11(1), 1-20 (2024)
7. Rao, S.U.M., Sreekala, K., Rao, P.S., Shirisha, N., Srinivas, G., Sreedevi, E.: Plant disease classification using novel integration of deep learning CNN and graph convolutional networks. Indonesian Journal of Electrical Engineering and Computer Science 36(3), 1721-1730 (2024)
8. Bera, A., Bhattacharjee, D., Krejcar, O.: PND-Net: Plant nutrition deficiency and disease classification using graph convolutional networks. Scientific Reports 14, 66543 (2024)
9. Maruthai, S., Selvanarayanan, R., Thanarajan, T., Rajendran, S.: Hybrid vision GNNs based early detection and protection against pest diseases in coffee plants. *Sci Rep* 15, 11778 (2025)
10. Woebbecke, D.M., Meyer, G.E., Von Bargen, K., Mortensen, D.A.: Color indices for weed identification under various soil, residue, and lighting conditions. Transactions of the ASAE 38(1), 259-269 (1995)
11. Yang, J., Xu, G., Yang, M., Lin, Z.: Lightweight wavelet-CNN tea leaf disease detection. PLoS ONE, 20(5), e0323322 (2025)
12. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC Superpixels Compared to State-of-the-art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11), 2274-2282 (2012)
13. Gascó, E., López-González, C.I., Besada-Portas, E., Pajares, G.: Hybrid Integration of CNN, Spectral, Structural, and Attention Mechanisms for Vegetation-Soil Segmentation in Agricultural Images. In: Corchado, E., et al. Hybrid Artificial Intelligent Systems. HAIS 2026. Lecture Notes in Computer Science, vol. 16596, pp. 134-145. Springer, Cham. (2026)
14. MathWorks Homepage, https://es.mathworks.com/, last accessed 2026/03/26
15. ETfos Research Group. Crop Row Benchmark Dataset. Homepage, http://www.etfos.unios.hr/r3dvgroup/index.php?id=crd_dataset, last accessed 2026/03/26