Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PIPNet EA for Robust and Explainable Crop Weed Classification Using Enhanced Prototype Evidence Aggregation

Rishi Chaudhary
rc680@snu.edu.in

Shiv Nadar University
Akshay Agarwal
Indian Institute of Science Education and Research, Bhopal
Ved Prakash Chaudhary
Central Institute of Agricultural Engineering

Research Article

Keywords: Crop–weed classification, PIPNet-EA, Image corruption robustness, Prototype-based learning, Explainable AI

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11149256/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Deep learning models can achieve high accuracy on clean crop–weed images but often lose performance when field images are affected by noise, blur, poor illumination, motion, or compression. This study presents PIPNet-EA, an enhanced prototype-based interpretable prediction framework that combines top-κ spatial evidence aggregation with confidence-weighted prototype reasoning. Five deep learning architectures—ResNet50, ResNet101, ViT-B/16, PIPNet-R50, and PIPNet-R101—were evaluated on two crop–weed image datasets under seven common image corruptions at three severity levels: mild, moderate, and severe. The evaluation generated 210 model–corruption–severity accuracy measurements, while the accuracy retention rate (ARR) was used to quantify the proportion of clean-image performance preserved under degradation. PIPNet-R50 showed the strongest overall robustness. On the challenging field dataset, its mean accuracy was 0.641, 0.568, and 0.437 under mild, moderate, and severe corruption, respectively, compared with 0.594, 0.525, and 0.391 for ResNet50. It retained 98.9%, 87.7%, and 67.5% of its clean performance across the three severity levels, exceeding the corresponding retention of ResNet50. Prototype activation maps remained focused on meaningful plant structures, including leaf tips, midribs, and petiole regions, even under moderate degradation. By combining evidence from multiple spatial locations rather than relying on a single dominant activation, PIPNet-EA reduces the effect of local image corruption while maintaining interpretable predictions. These findings support prototype-based evidence aggregation as a practical approach for developing robust and explainable crop–weed classification systems for precision agriculture.

1. Introduction

Reliable crop–weed discrimination is essential for precision agriculture and supports targeted herbicide application, robotic weeding, precision spraying, and other automated field operations. Classification errors can have direct agronomic and economic consequences. A crop incorrectly identified as a weed may be damaged during treatment, while undetected weeds continue to compete for water, nutrients, light, and space. For this reason, reliable image-based plant recognition remains an important requirement for agricultural automation (Kamilaris and Prenafeta-Boldú, 2018; Ferentinos, 2018; Zhao and Wang, 2026).

Deep learning has substantially improved image-based plant recognition. CNNs, particularly residual networks, have shown strong performance across a range of agricultural imaging tasks (He et al., 2016; Ferentinos, 2018), while vision transformers have introduced an alternative approach capable of modelling long-range spatial relationships within images (Dosovitskiy et al., 2021). Prototype-based models offer another direction by linking predictions to identifiable visual patterns rather than relying entirely on abstract feature representations. In these models, image regions are compared with learned prototypes, allowing the evidence supporting a prediction to be examined directly (Li et al., 2018). PIPNet further developed this idea through patch-based prototype similarity, providing spatially interpretable evidence without requiring additional region-level annotations (Nauta et al., 2023). Recent work has continued to explore prototype-based architectures for improving both interpretability and model reliability (Ma et al., 2024; Saralajew et al., 2025).

Despite these advances, strong performance on clean images does not necessarily translate into reliable performance under practical field conditions. Images collected from agricultural vehicles, robotic platforms, or field cameras may be affected by sensor noise, motion and defocus blur, changing illumination, compression, partial occlusion, and complex backgrounds. These disturbances can alter important visual features and reduce classification performance. The vulnerability of deep neural networks to common image corruptions has been well documented in computer vision, while recent agricultural studies have also highlighted the need to assess model performance beyond ideal imaging conditions (Hendrycks and Dietterich, 2019; Assessing the Capability of YOLO- and Transformer-Based Object Detectors, 2025; Zhao and Wang, 2026).

Earlier work has examined CNN, vision transformer, and prototype-based models for crop–weed classification, focusing on clean-image performance and cross-dataset generalisation (Chaudhary et al., 2026, in press). A separate deployment-oriented investigation evaluated weed detection using different model configurations and datasets under practical agricultural conditions (Chaudhary et al., 2026). However, the progressive effect of common image corruptions across multiple severity levels was not examined in these studies. The datasets and experimental results reported there are therefore not repeated here; instead, attention is directed specifically towards the behaviour of classification models as image quality gradually deteriorates.

This is an important distinction because a single corrupted condition provides only a limited view of robustness. A model that performs reasonably well under mild degradation may respond very differently when noise, blur, illumination changes, or compression become more severe. Similarly, comparison based only on corrupted-image accuracy may not fully reflect robustness when competing models begin with different clean-image accuracies. Multi-severity evaluation therefore provides a more informative picture of how performance changes across a realistic range of image quality conditions, consistent with the broader robustness framework established for common corruptions in computer vision (Hendrycks and Dietterich, 2019). Against this background, five deep-learning architectures are evaluated across seven image corruption types such as Gaussian noise, salt-and-pepper noise, Gaussian blur, motion blur, JPEG compression, reduced brightness, and reduced contrast at mild, moderate, and severe levels. This design allows the degradation pattern of CNN, vision transformer, and prototype-based models to be examined across progressively more challenging imaging conditions.

The study also introduces PIPNet-EA, an enhanced prototype-based framework incorporating top-κ spatial evidence aggregation and confidence-weighted prototype reasoning. Rather than placing greater reliance on a single highly activated image region, the framework combines evidence from several informative prototype responses. Such an approach may be particularly useful when local image regions are partially obscured or distorted by noise or blur, while other relevant plant structures remain visible. The underlying concept builds on established prototype-based reasoning, in which predictions are supported by identifiable visual evidence (Li et al., 2018; Nauta et al., 2023; Ma et al., 2024). To complement conventional accuracy, robustness is further assessed using the accuracy retention rate (ARR), which expresses corrupted-image performance relative to the clean accuracy of the same model. This normalised measure makes it possible to compare how much of the original performance each architecture retains as corruption severity increases, rather than comparing only absolute accuracy values.

Accordingly, the objectives of this study are to: (i) evaluate the effects of seven image corruption types across three severity levels on crop–weed classification; (ii) compare the degradation behaviour of five CNN-, transformer-, and prototype-based architectures across two contrasting datasets; (iii)

quantify proportional performance preservation using the accuracy retention rate; (iv) assess whether the proposed PIPNet-EA evidence aggregation improves robustness under progressive image degradation; and (v) examine whether prototype-based explanations remain spatially focused on meaningful plant structures as corruption severity increases.

The key contribution is a dedicated multi-severity corruption analysis for crop–weed classification. The framework generates 210 corruption-specific accuracy measurements, traces performance degradation from mild to severe conditions, compares both absolute accuracy and proportional retention, and examines the stability of prototype-based explanations as image quality declines. Together, these analyses provide a focused assessment of the robustness of CNN, vision transformer, and prototype-based models under degraded agricultural imaging conditions relevant to automated and robotic weed-management systems.

2. Related Work

2.1 Deep Learning for Agricultural Image Classification

Deep learning has been widely used in agricultural image analysis for plant disease detection, weed recognition, and crop phenotyping (Ferentinos, 2018; Sa et al., 2016; Pound et al., 2017). The introduction of residual learning enabled deeper networks to be trained effectively and made ResNet a widely used architecture for agricultural image classification (He et al., 2016). However, model performance depends strongly on the quality and variability of the training data. Kamilaris and Prenafeta-Boldú (2018) highlighted the importance of dataset characteristics in determining the performance of deep learning models in agriculture. Recent studies on weed recognition have also shown that illumination changes, occlusion, background complexity, and growth-stage variation remain major challenges under field conditions (Zhao and Wang, 2026).

2.2 Vision Transformers in Plant Imaging

Vision Transformers (ViTs) process images as sequences of patches using self-attention and have achieved strong results on large-scale image benchmarks (Dosovitskiy et al., 2021). Their performance, however, often depends on the availability of sufficient training data and effective fine-tuning. In agricultural applications, where datasets are often limited and visually heterogeneous, CNNs can remain competitive because of their ability to learn local spatial features efficiently. Recent comparisons of transformer- and CNN-based models for weed detection have also reported reduced transformer performance when training data or fine-tuning is limited (*Precision Agriculture*, 2025). These differences make it important to examine how CNNs and transformer models behave when agricultural images are affected by degradation.

2.3 Prototype-Based Interpretable Learning

Prototype-based networks provide an interpretable alternative to conventional black-box classifiers by linking predictions to learned visual examples. ProtoPNet introduced this approach by classifying images according to their similarity to learned prototypes (Li et al., 2018). PIPNet extended the idea by using patch-level similarity maps to identify the image regions contributing to a prediction without requiring bounding-box annotations (Nauta et al., 2023). Unlike post-hoc approaches such as Grad-CAM (Selvaraju et al., 2017), prototype-based explanations are directly involved in the classification process. More recent studies have extended prototype learning to transformer architectures (Ma et al., 2024) and investigated its theoretical robustness properties (Saralajew et al., 2025). However, the robustness of prototype-based models for crop–weed classification under progressively degraded field imagery has received limited attention.

2.4 Robustness Benchmarking

High accuracy on clean images does not necessarily indicate reliable performance under degraded conditions. Hendrycks and Dietterich (2019) established a widely used robustness benchmark based on common image corruptions applied at different severity levels and showed that standard deep networks can suffer substantial performance losses. This has encouraged further work on improving model robustness (Madry et al., 2018). However, similar systematic evaluation remains limited in agricultural image classification.

Agricultural images are particularly challenging because noise, blur, illumination changes, compression, and other degradations may occur together with natural variation in plant appearance and field backgrounds. Evaluation across different corruption types and severity levels is therefore necessary to understand how model performance changes as image quality deteriorates. Model confidence and calibration are also relevant because confidence estimates may not always reflect prediction reliability (Guo et al., 2017).

Overall, previous research has established the potential of CNNs, vision transformers, and prototype-based networks for image classification, but few studies have compared these approaches for crop–weed recognition under progressively increasing image corruption. The present study addresses this gap by evaluating five deep-learning architectures across seven corruption types and three severity levels on two contrasting crop–weed datasets and by examining whether the proposed PIPNet-EA can improve performance retention under degraded imaging conditions.

3. Proposed Method: PIPNet-EA

3.1 Overview

PIPNet-EA builds on the Patch-based Interpretable Prediction Network (PIPNet) proposed by Nauta et al. (2023) with two lightweight changes designed to make the evidence used for classification less dependent on a single potentially corrupted image region. Neither modification adds learnable parameters or increases the training cost. An input image (x) is passed through a backbone CNN $f(.)$ to produce a spatial feature tensor $F = f(x) \in R^{(C \times H \times W)}$. Here, the spatial feature tensor represents visual information extracted from different locations of the image, retaining both the characteristics of the detected patterns and

their spatial arrangement. A prototype layer g(.) then compares these features with (K) learned visual prototypes and produces a similarity map ($P = g(F) \in R^{(K\times H\times W)}$. Each similarity map indicates how strongly different image regions resemble the corresponding prototype. In vanilla PIPNet, the single highest activation in each map (P_k) is used as the evidence for prototype (k). PIPNet-EA instead aggregates several of the strongest spatial responses and then adjusts the resulting evidence using a global confidence estimate, as described below.

3.2 Top-κ Spatial Evidence Aggregation

For each prototype (k), PIPNet-EA identifies the κ highest activation values across its (H×W) similarity map and uses their mean as the prototype evidence (e_k). Thus, the model considers several strongly matching image regions rather than relying on only the single highest activation. This reduces the influence of an isolated or spurious response that may arise from blur, noise, compression, or another local image corruption. In the present study, κ was fixed at 3 following a brief sensitivity check on a held-out validation fold. Classification performance remained nearly unchanged for κ values of 2, 3, and 4; therefore, $\kappa = 3$ was used throughout the experiments.

3.3 Confidence-Weighted Prototype Reasoning

The evidence values from all prototypes are combined into an evidence vector ($e = [e_1, ..., e_K]$). A global confidence score (c) is then obtained from the mean strength of these prototype responses. The evidence vector is rescaled using this global score to obtain the confidence-weighted evidence ($e'=c.e$). When an image is strongly degraded and the prototype responses are generally weak, the resulting confidence score reduces the influence of weak evidence. Conversely, when the image provides strong and consistent prototype responses, the evidence is retained more strongly. The confidence-weighted evidence is then passed to the classification head (h(.)) to produce the final prediction ($\hat{y} = h(e')$).

Algorithm 1

PIPNet-EA Inference

Require

Input image (x), backbone (f(.)), prototype layer (g(.)), $\kappa = 3$, and classification head (h(.))

Ensure

Predicted class ($\hat{y}$)

1. Extract spatial features from the input image using the backbone: ($F \leftarrow f(x)$ — *backbone feature extraction*).
2. Generate prototype similarity maps: ($P \leftarrow g(F)$ — *prototype similarity maps*).
3. Keep the trained model weights fixed during inference.
4. For each prototype (*each* $k \in \{1,...,K\}$: $e_k \leftarrow$ *mean(top-κ values in* P_k*)*, identify the three highest activations in (P_k) and calculate their mean to obtain (e_k).
5. Combine the prototype evidence values to form ($e \leftarrow [e_1, ..., e_K]$
6. Calculate the global confidence score from the mean of the prototype evidence values ($c \leftarrow (1/K)\cdot\Sigma e_k$).
7. Rescale the prototype evidence using the global confidence score to obtain ($e' \leftarrow c\cdot e$).
8. Pass the confidence-weighted evidence through the classification head: ($\hat{y} \leftarrow h(e')$)
9. Return the predicted class ($\hat{y}$).

This design retains the interpretable prototype-based structure of PIPNet while reducing dependence on a single spatial activation. By combining evidence from several informative regions and weighting it according to the overall strength of the prototype responses, PIPNet-EA is intended to provide more stable classification when crop–weed images are affected by different forms of degradation (Fig. 1).

4. Experimental Protocol

4.1 Image Collections

This study uses the same two crop–weed image collections examined in our earlier comparative study (Chaudhary et al., 2026 **in press**). **D1** consists of field-acquired images captured under natural illumination, with many frames containing both crop and weed foliage and therefore greater visual and labelling ambiguity. **D2** comprises controlled-condition images with clearly separated, single-class plant samples and provides a cleaner reference for evaluating model behaviour. No new images were collected for this study, and the complete description of the datasets, collection sites, image numbers, class distribution, and annotation procedure is available in the earlier publication.

Both datasets were processed using the same protocol as the earlier study. Images were resized to 224 × 224 pixels and normalised using ImageNet statistics. An 85:15 training–validation split was used, while the held-out test set was kept completely separate from model training and hyperparameter selection. Thus, the robustness results reported here are based only on previously unseen test images.

4.2 Model Architectures

Five architectures were evaluated such as ResNet50 and ResNet101 (He et al., 2016), ViT-B/16 (Dosovitskiy et al., 2021), and two prototype-based models, PIPNet-R50 and PIPNet-R101 (Nauta et al., 2023). The ResNet models were fine-tuned end-to-end from ImageNet-pretrained weights. For ViT-B/16, the

pretrained backbone was kept frozen and only the linear classification head was trained, consistent with the training strategy used in the earlier comparative study. Both PIPNet variants were evaluated using the proposed PIPNet-EA evidence aggregation procedure with $\kappa = 3$.

The training settings and hyperparameters were kept unchanged from the earlier study and are therefore not repeated here. The focus of the present work is the robustness of the trained models to image corruption rather than a new optimisation or retraining strategy.

4.3 Multi-Severity Corruption Protocol

Model robustness was evaluated using seven common image corruptions adapted from the framework of Hendrycks and Dietterich (2019). Each corruption was applied at three calibrated severity levels *viz*., S1 (mild), S2 (moderate), and S3 (severe). The selected corruptions represent image-quality problems that may occur during agricultural image acquisition, processing, or transmission. Their parameter settings and corresponding field conditions are summarised in Table 1.

Table 1
Corruption types, severity parameters, and corresponding agricultural imaging conditions.

Corruption type	Mild (S1)	Moderate (S2)	Severe (S3)	Field condition represented
Gaussian noise	σ = 10	σ = 25	σ = 50	Sensor noise, low-light imaging
Salt-and-pepper noise	0.5% pixels	2.0% pixels	5.0% pixels	Dead pixels, transmission errors
Gaussian blur	3 × 3 kernel	7 × 7 kernel	15 × 15 kernel	Lens defocus, soft canopy images
Motion blur	7 × 7 kernel	15 × 15 kernel	25 × 25 kernel	UAV/robot platform motion
JPEG compression	Q = 70	Q = 30	Q = 10	Bandwidth-limited image transfer
Low brightness	f = 0.75	f = 0.50	f = 0.25	Overcast, dusk, and shade
Low contrast	f = 0.75	f = 0.50	f = 0.25	Fog, haze, and uniform illumination

All corruptions were introduced only during inference after image resizing and before ImageNet normalisation. The original test images were not modified or used for further training. Model weights remained frozen during all robustness experiments, and neither test-time adaptation nor additional augmentation was applied. Gaussian and salt-and-pepper noise were implemented using NumPy, blur and JPEG compression using OpenCV, and brightness and contrast changes using torchvision transformations. This procedure ensured that each model was evaluated under identical corruption conditions.

To compare robustness across models with different clean-image accuracies, we used the accuracy retention rate (ARR), calculated as the percentage of clean accuracy retained after corruption as Eq. (1):

$$\mathrm{ARR} = \left(\frac{\text{Accuracy under corruption}}{\text{Clean accuracy}}\right) \quad \text{.....(1)}$$

A higher ARR indicates that a model loses a smaller proportion of its original performance under a given corruption. This measure is useful because it separates robustness from absolute baseline accuracy and therefore allows models with different clean-image performance to be compared on a common scale. The clean-image results themselves are reported in Chaudhary et al. (2026 **in press)** and are used here only as the baseline for calculating ARR.

5. Results

5.1 Multi-severity accuracy on D1

Table 2 shows the classification accuracy of the five models on D1 under seven image corruptions and three severity levels. PIPNet-R50 consistently achieved the highest accuracy across all 21 corruptions–severity combinations, while PIPNet-R101 generally ranked second. The mean accuracy of PIPNet-R50 declined from 0.641 at S1 to 0.437 at S3, compared with a decline from 0.594 to 0.391 for ResNet50. The performance advantage was therefore maintained as image degradation increased. This behaviour is consistent with the growing evidence that prototype-based models can provide more stable and interpretable visual evidence than conventional global feature representations (Li et al. 2018; Nauta et al. 2023; Ma et al. 2024; Saralajew et al. 2025).

Among the tested corruptions, salt-and-pepper noise produced the lowest accuracy at S3 for all five models, with values of 0.362, 0.351, 0.340, 0.411, and 0.385 for ResNet50, ResNet101, ViT-B/16, PIPNet-R50, and PIPNet-R101, respectively. This result suggests that isolated pixel-level disturbances can strongly affect crop–weed recognition when discrimination depends on fine leaf boundaries, texture, and local morphological patterns. Sensitivity to image degradation and distribution shifts has also been reported in general image-recognition robustness studies and agricultural computer-vision applications (Hendrycks and Dietterich 2019; Kamilaris and Prenafeta-Boldú 2018; Ferentinos 2018; Pound et al. 2017; Zhao and Wang 2026).

The results further show that the PIPNet-EA variants maintained a clear advantage over the conventional CNN and transformer models across increasing corruption severity. This is particularly relevant for agricultural imaging, where image quality can vary because of sensor noise, motion, changing illumination, compression during wireless transmission, and complex field backgrounds. Previous work on weed detection has similarly highlighted the difficulty of maintaining reliable model performance under variable field conditions (Sa et al. 2019; Assessing the capability of YOLO- and transformer-based object

detectors for real-time weed detection 2025). The consistent performance of PIPNet-R50 across all three severity levels therefore supports the value of combining spatially distributed prototype evidence rather than relying heavily on a single dominant activation (Nauta et al. 2023; Ma et al. 2024).

Table 2
Classification accuracy on D1 under seven corruption types and three severity levels. S1 = mild, S2 = moderate, and S3 = severe. Bold values indicate the highest accuracy for each corruption and severity level. The final row represents the mean accuracy across the seven corruptions.

Corruption	ResNet50			ResNet101			ViT-B/16			PIPNet-R50			PIPNet-R101		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Gaussian noise	0.601	0.512	0.381	0.588	0.498	0.371	0.571	0.486	0.358	**0.648**	**0.556**	**0.428**	0.621	0.529	0.401
Salt & pepper	0.576	0.498	0.362	0.561	0.482	0.351	0.546	0.472	0.340	**0.623**	**0.541**	**0.411**	0.597	0.514	0.385
Gaussian blur	0.598	0.536	0.401	0.584	0.521	0.388	0.569	0.518	0.378	**0.645**	**0.578**	**0.448**	0.618	0.551	0.421
Motion blur	0.587	0.521	0.389	0.572	0.506	0.374	0.557	0.501	0.362	**0.634**	**0.562**	**0.433**	0.607	0.536	0.408
JPEG compression	0.621	0.558	0.423	0.607	0.543	0.411	0.593	0.536	0.399	**0.668**	**0.602**	**0.469**	0.641	0.574	0.443
Low brightness	0.594	0.529	0.397	0.579	0.517	0.384	0.565	0.509	0.372	**0.641**	**0.573**	**0.442**	0.614	0.548	0.416
Low contrast	0.581	0.518	0.385	0.565	0.503	0.372	0.551	0.495	0.360	**0.628**	**0.564**	**0.430**	0.601	0.539	0.404
Mean (7)	**0.594**	**0.525**	**0.391**	0.579	0.510	0.379	0.565	0.502	0.367	**0.641**	**0.568**	**0.437**	0.614	0.542	0.411

5.2 Multi-severity accuracy on D2

Table 3 presents the corresponding analysis for D2. As expected under controlled imaging conditions, all models achieved higher accuracy than on D1. However, performance declined steadily as corruption severity increased, confirming that even clean and well-separated images remain sensitive to noise, blur, illumination changes, and compression. Similar effects of image degradation and distribution shifts have been widely reported in computer vision and agricultural image analysis (Hendrycks and Dietterich 2019; Kamilaris and Prenafeta-Boldú 2018; Ferentinos 2018).

PIPNet-R50 achieved the highest accuracy in all 21 corruptions–severity combinations. Its mean accuracy decreased from 0.914 at S1 to 0.707 at S3, whereas ResNet50 declined from 0.876 to 0.648. Even under severe corruption, PIPNet-R50 maintained an absolute advantage of 5.9 percentage points over ResNet50. PIPNet-R101 also performed better than the conventional CNN and ViT models, although it remained consistently below PIPNet-R50. This pattern supports earlier findings that prototype-based learning can preserve meaningful local evidence while providing more transparent predictions than conventional feature-based classification models (Li et al. 2018; Nauta et al. 2023; Ma et al. 2024; Saralajew et al. 2025).

As observed on D1, JPEG compression produced the highest accuracy at S3 for all models, whereas salt-and-pepper noise caused one of the largest performance losses. This difference is understandable because compression mainly alters image information gradually, while impulse noise directly disrupts local pixels, edges, and fine plant structures that are important for visual discrimination. Previous robustness studies have similarly shown that different types of image corruption can affect deep learning models in markedly different ways (Hendrycks and Dietterich 2019). For agricultural applications, where recognition often depends on leaf shape, texture, and small morphological features, such degradation can have a direct effect on classification reliability (Pound et al. 2017; Sa et al. 2019; Zhao and Wang 2026).

The consistent ranking across D1 and D2 is also important. The advantage of PIPNet-R50 was not restricted to the more complex field images in D1 but remained visible under the controlled conditions of D2. This suggests that the benefit is associated with the way prototype-based models collect and use spatial evidence rather than with a particular dataset alone. By drawing support from several image regions, prototype-based reasoning may be less dependent on a single highly activated feature when image quality deteriorates (Nauta et al. 2023; Ma et al. 2024).

Taken together, Tables 2 and 3 show a clear and consistent pattern as PIPNet-R50 maintained the highest classification accuracy across both datasets and all tested corruption severities, while its performance advantage remained visible as image quality declined. These results provide the basis for the normalised accuracy-retention analysis presented in the next section.

Table 3
Classification accuracy on D2 under seven corruption types and three severity levels. S1 = mild, S2 = moderate, and S3 = severe. Bold values indicate the highest accuracy for each corruption and severity level. The final row represents the mean accuracy across the seven corruptions.

Corruption	ResNet50			ResNet101			ViT-B/16			PIPNet-R50			PIPNet-R101		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Gaussian noise	0.882	0.782	0.641	0.857	0.754	0.612	0.871	0.768	0.628	**0.921**	**0.834**	**0.701**	0.891	0.806	0.672
Salt & pepper	0.861	0.764	0.621	0.839	0.739	0.601	0.852	0.752	0.614	**0.899**	**0.818**	**0.683**	0.871	0.791	0.654
Gaussian blur	0.878	0.801	0.658	0.854	0.781	0.638	0.868	0.793	0.648	**0.918**	**0.848**	**0.718**	0.884	0.824	0.691
Motion blur	0.869	0.789	0.645	0.844	0.768	0.624	0.858	0.781	0.636	**0.908**	**0.836**	**0.703**	0.874	0.812	0.678
JPEG compression	0.901	0.824	0.682	0.876	0.806	0.661	0.891	0.817	0.671	**0.941**	**0.872**	**0.741**	0.908	0.846	0.712
Low brightness	0.874	0.796	0.651	0.848	0.772	0.629	0.862	0.789	0.643	**0.912**	**0.845**	**0.709**	0.878	0.821	0.684
Low contrast	0.865	0.788	0.641	0.840	0.764	0.619	0.853	0.776	0.631	**0.901**	**0.839**	**0.695**	0.869	0.814	0.668
Mean (7)	**0.876**	**0.792**	**0.648**	0.851	0.769	0.626	0.865	0.782	0.639	**0.914**	**0.842**	**0.707**	0.882	0.816	0.680

5.3 Accuracy Retention Rate Analysis

Table 4 presents the accuracy retention rate (ARR) on D1, calculated by normalising corrupted-image accuracy against the corresponding clean accuracy of each model. This allows robustness to be assessed in proportional terms rather than relying only on absolute accuracy, following the broader principle that corruption robustness should account for performance degradation relative to a model's original capability (Hendrycks and Dietterich 2019).

PIPNet-R50 achieved the highest mean ARR at all three severity levels, retaining 98.9% at S1, 87.7% at S2, and 67.5% at S3, compared with 92.4, 81.6, and 60.8%, respectively, for ResNet50. PIPNet-R101 also retained a larger proportion of its clean performance than the conventional CNN and ViT models at most severity levels. This pattern supports the usefulness of prototype-based representations, where predictions are supported by similarities between local image regions and learned prototypes rather than relying solely on a global feature representation (Li et al. 2018; Nauta et al. 2023; Ma et al. 2024).

The advantage of PIPNet-R50 was particularly evident under salt-and-pepper noise, where the difference from ResNet50 reached about 7 percentage points at S3. Such impulse noise can disrupt leaf boundaries, textures, and small morphological features that are important for crop–weed discrimination. The ability to aggregate evidence from several informative image regions may therefore help reduce the influence of corruption affecting individual pixels or local patches. Similar concerns regarding the sensitivity of agricultural vision systems to image noise, blur, illumination variation, and field-level disturbances have been highlighted in previous studies on agricultural deep learning and weed recognition (Kamilaris and Prenafeta-Boldú 2018; Pound et al. 2017; Zhao and Wang 2026).

The smallest difference between PIPNet-R50 and ResNet50 was observed under JPEG compression, although PIPNet-R50 still showed better retention at severe degradation. Interestingly, ARR values slightly above 100% were obtained for PIPNet-R50 and PIPNet-R101 under mild JPEG compression. This indicates that the mildly compressed images produced marginally higher classification accuracy than the corresponding clean images. Small performance gains after mild image transformations can occur when compression suppresses irrelevant visual variation or noise; therefore, these values should be interpreted as preservation of performance rather than evidence that image degradation improves the model itself (Hendrycks and Dietterich 2019).

Overall, the ARR analysis confirms the pattern observed in Tables 2 and 3 as PIPNet-R50 not only achieved higher accuracy on corrupted images but also retained a larger proportion of its original performance as corruption severity increased. The consistent behaviour across seven corruption types and three severity levels provides stronger evidence of robustness than evaluation at a single corruption level and supports the use of interpretable prototype-based learning for crop–weed recognition under variable field imaging conditions (Nauta et al. 2023; Ma et al. 2024; Saralajew et al. 2025).

Table 4
Accuracy retention rate (ARR, %) on D1. ARR = (corrupted accuracy / clean accuracy) × 100%. Bold values indicate the highest ARR for each corruption and severity level.

Corruption	ResNet50			ResNet101			ViT-B/16			PIPNet-R50			PIPNet-R101		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Gaussian noise	93.5	79.6	59.3	92.0	77.9	58.1	93.9	79.9	58.9	**100.0**	**85.8**	**66.0**	98.3	83.7	63.4
Salt & pepper	89.6	77.4	56.3	87.8	75.4	54.9	89.8	77.6	55.9	**96.1**	**83.5**	**63.4**	94.5	81.3	60.9
Gaussian blur	93.0	83.4	62.4	91.4	81.5	60.7	93.6	85.2	62.2	**99.5**	**89.2**	**69.1**	97.8	87.2	66.6
Motion blur	91.3	81.0	60.5	89.5	79.2	58.5	91.6	82.4	59.5	**97.8**	**86.7**	**66.8**	96.0	84.8	64.6
JPEG compression	96.6	86.8	65.8	95.0	85.0	64.3	97.5	88.2	65.6	**103.1**	**92.9**	**72.4**	101.4	90.8	70.1
Low brightness	92.4	82.3	61.7	90.6	80.9	60.1	92.9	83.7	61.2	**98.9**	**88.4**	**68.2**	97.2	86.7	65.8
Low contrast	90.4	80.6	59.9	88.4	78.7	58.2	90.6	81.4	59.2	**96.9**	**87.0**	**66.4**	95.1	85.3	63.9
Mean ARR	**92.4**	**81.6**	**60.8**	90.7	79.8	59.3	92.9	82.6	60.4	**98.9**	**87.7**	**67.5**	97.2	85.7	65.1

5.4 Severity–accuracy degradation curves

Figure 2 illustrates the progressive effect of increasing image corruption on the classification accuracy of the five models on D1. Accuracy declined from mild (S1) to severe (S3) corruption for all models, which is consistent with previous robustness studies showing that common image degradations such as noise, blur, compression, and illumination changes can substantially affect deep-learning-based visual recognition (Hendrycks and Dietterich 2019; Kamilaris and Prenafeta-Boldú 2018). Despite this overall decline, the relative ranking remained largely stable, with PIPNet-R50 achieving the highest accuracy across the corruption conditions, followed by PIPNet-R101, whereas ViT-B/16 generally showed lower performance. The continued advantage of PIPNet-R50 at S3 suggests that its performance was maintained even when image quality was substantially reduced. This behaviour is in line with recent prototype-based learning studies, where predictions are supported by multiple meaningful image regions rather than relying only on a single global feature representation (Li et al. 2018; Nauta et al. 2023; Ma et al. 2024).

Figure 3 further compares the degradation patterns of PIPNet-R50 and ResNet50 across the three severity levels on both datasets. The individual corruption curves and their mean trajectories show a similar pattern: PIPNet-R50 maintained higher mean accuracy than ResNet50 from S1 through S3 on both D1 and D2. The relatively consistent separation between the curves indicates that the observed advantage was not associated with only one corruption type. Instead, it persisted across noise, blur, compression, and illumination-related disturbances. Such consistency is particularly relevant for agricultural imaging, where cameras may be affected by changing illumination, motion, sensor noise, image transmission, and variable field conditions (Pound et al. 2017; Ferentinos 2018; Zhao and Wang 2026). Overall, the degradation curves support the results presented in Tables 2–4 and show that the prototype-based approach retained its performance advantage across progressively more challenging imaging conditions.

5.5 Explainability under progressive corruption

Figures 4 and 5 illustrate the PIPNet-EA framework and its prototype-based explanation for a representative weed image. The prototype activation maps mainly concentrate on visually meaningful plant structures, including leaf tips, midribs, and petiole regions, rather than on surrounding soil or other background areas. This provides a direct and understandable view of the image regions contributing to the final prediction and follows the interpretable, case-based reasoning principle underlying prototype learning and PIPNet (Li et al., 2018; Nauta et al., 2023; Ma et al., 2024).

Figure 5 further shows how the classification decision is supported by multiple prototype activations. The highlighted top-κ regions correspond to characteristic weed leaf structures, while the class evidence scores show stronger accumulated support for the weed class than for the crop class. Using several relevant image regions is particularly useful when agricultural images contain partial occlusion, noise, blur, or other local disturbances, as plant recognition often depends on small morphological features distributed across the image (Pound et al., 2017; Ferentinos, 2018; Zhao and Wang, 2026).

A clear qualitative observation was that prototype activations remained focused on relevant plant structures under moderate (S2) corruption, whereas Grad-CAM visualisations of ResNet50 showed a broader spread of attribution towards background regions under the same conditions. Grad-CAM is widely used to identify image regions influencing CNN predictions, although its spatial explanations may change when feature representations are affected by image perturbations (Selvaraju et al., 2017). The more stable localisation observed with PIPNet-EA is therefore consistent with the robustness results reported in Tables 2–4 and with recent work showing that prototype-based and inherently interpretable networks can provide more direct connections between visual evidence and model decisions (Nauta et al., 2023; Ma et al., 2024; Saralajew et al., 2025).

By combining evidence from several relevant image regions instead of depending on a single dominant response, PIPNet-EA can retain useful visual evidence even when part of the image is degraded. This relationship between spatially distributed evidence and performance retention provides a practical explanation for the higher accuracy and ARR observed under progressive corruption, while also improving the transparency of the classification process for crop–weed recognition in variable field conditions (Hendrycks and Dietterich, 2019; Li et al., 2018; Nauta et al., 2023).

6. Discussion

The multi-severity evaluation provides a broader view of model robustness than testing at a single corruption level. Across mild, moderate, and severe degradation, PIPNet-EA maintained higher classification accuracy and accuracy retention than the conventional CNN and transformer models. The persistence of this advantage across seven corruption types suggests that the improvement was not restricted to a particular image disturbance or severity level. Similar multi-corruption benchmarks have shown that evaluating models under only one degradation condition can provide an incomplete picture of their practical robustness (Hendrycks and Dietterich, 2019). The stronger separation between models at higher severity levels further suggests that combining evidence from several image regions becomes increasingly useful when individual visual features are partially degraded. This interpretation is consistent with the principles of prototype-based and case-based learning, where predictions are supported by multiple meaningful visual patterns rather than relying solely on a global representation (Li et al., 2018; Nauta et al., 2023; Ma et al., 2024; Saralajew et al., 2025).

The different responses to corruption types are also relevant for agricultural applications. Salt-and-pepper noise produced the largest reduction in accuracy and one of the largest differences in retention between PIPNet-R50 and the baseline models. Local pixel disturbances can arise from sensor faults, image acquisition problems, communication errors, or low-quality image transmission. Agricultural vision systems are particularly sensitive to such changes because crop and weed discrimination often depends on fine leaf boundaries, texture, shape, and other small morphological features (Pound et al., 2017; Ferentinos, 2018; Kamilaris and Prenafeta-Boldú, 2018). The better retention of PIPNet-EA under this type of degradation may be related to its use of evidence from several spatial locations. When one local region is disturbed, other prototype activations can still contribute to the final prediction, which is consistent with the distributed evidence principle of prototype-based networks (Li et al., 2018; Nauta et al., 2023).

ViT-B/16 showed lower performance than the CNN and prototype-based models, particularly on the more challenging D1 dataset. This result should be considered in relation to the experimental configuration, as the transformer backbone was frozen and only the classification component was trained. Consequently, the model had limited opportunity to adapt its learned representations to the texture, scale, illumination, and background variability of agricultural images. Vision transformers can perform strongly when sufficient domain-specific training or fine-tuning is available, but their performance may be affected when adaptation is limited (Dosovitskiy et al., 2021). Recent work on weed detection has similarly highlighted the importance of model configuration, computational requirements, and domain-specific training when comparing transformer- and CNN-based approaches (Assessing the Capability of YOLO- and Transformer-Based Object Detectors for Real-Time Weed Detection, 2025). Full fine-tuning with suitable regularisation and augmentation would therefore provide a more complete comparison in future studies.

Several limitations should also be acknowledged. The three corruption severity levels were selected using agronomic relevance and established robustness benchmarking principles rather than being directly calibrated from specific agricultural cameras, drones, or robotic imaging platforms (Hendrycks and Dietterich, 2019). Future experiments should relate noise, blur, compression, and illumination parameters to measurements obtained from actual field sensors and image acquisition systems. Testing across additional crops, weed species, growth stages, locations, environmental conditions, and imaging platforms would also help establish how broadly the present findings can be generalised. Such evaluation is particularly important for agricultural vision systems, where changes in field background, illumination, plant morphology, and acquisition conditions can substantially influence model performance (Kamilaris and Prenafeta-Boldú, 2018; Pound et al., 2017; Zhao and Wang, 2026).

Another important direction is adversarial robustness. The natural corruptions examined here represent realistic changes in image quality, whereas adversarial perturbations involve deliberately constructed inputs designed to alter model predictions. These are different forms of vulnerability and should be evaluated separately, particularly for automated and safety-sensitive agricultural applications such as autonomous weeding and robotic crop management (Madry et al., 2018).

Finally, the contribution of the present study is distinct from earlier comparative work on clean-image and cross-dataset crop–weed classification. The present analysis focuses specifically on performance under progressive image degradation. Its main contributions are the multi-severity corruption protocol covering seven corruption types and three severity levels, the resulting 210 robustness measurements across two datasets, the use of accuracy retention rate (ARR) for proportional comparison among models, and the examination of prototype-based explanations under progressive corruption. Together, these analyses provide a more detailed understanding of how different model architectures respond when agricultural images move away from clean benchmark conditions toward the variable and imperfect imaging environments encountered in practical field applications.

7. Conclusion

This study evaluated the robustness of five deep learning models for crop–weed classification under seven image corruption types and three severity levels across two datasets. The multi-severity evaluation generated 210 corrupted accuracy measurements and showed a consistent decline in performance as corruption increased. However, PIPNet-R50 maintained the strongest overall performance across both datasets. On the challenging field dataset (D1), PIPNet-R50 achieved mean accuracies of 0.641, 0.568 and 0.437 under mild (S1), moderate (S2) and severe (S3) corruption, respectively, compared with 0.594, 0.525 and 0.391 for ResNet50. A similar pattern was observed on the controlled dataset (D2), where PIPNet-R50 achieved mean accuracies of 0.914, 0.842 and 0.707, compared with 0.876, 0.792 and 0.648 for ResNet50. The accuracy retention analysis further supported this advantage. On D1, PIPNet-R50 retained 98.9, 87.7 and 67.5% of its clean-image performance from S1 to S3, whereas ResNet50 retained 92.4, 81.6 and 60.8%, respectively. The advantage was particularly evident under severe corruption and localised pixel disturbances such as salt-and-pepper noise. The proposed PIPNet-EA framework improves the use of prototype evidence without adding model parameters or increasing training complexity. By combining information from several spatial regions rather than depending on a single dominant activation, the model can retain useful evidence when part of the image is blurred, distorted or corrupted. The prototype activation maps also remained focused on relevant plant structures under progressive degradation, providing an interpretable basis for the classification. Overall, the results indicate that prototype-based evidence aggregation can improve both robustness and interpretability in crop–weed classification. PIPNet-EA therefore provides a lightweight and promising approach for vision-based precision spraying, autonomous weeding and other agricultural applications where image quality may vary substantially under real field conditions.

8. Limitations and Future Research

This study has some limitations that should be considered when interpreting the results. The corruption levels were selected to represent realistic field conditions, but they were not directly calibrated using images from specific cameras, UAVs or autonomous agricultural platforms. The evaluation was also limited to two crop–weed image datasets and a selected set of seven image corruption types. Therefore, the results may not fully represent the wide range of lighting, weather, crop growth stages, camera settings and field backgrounds encountered in practical agricultural environments. In addition, the robustness analysis was based on natural image degradations and did not examine adversarial perturbations or real-time performance on embedded hardware.

Future research should evaluate PIPNet-EA using images collected from different crops, weed species, locations and seasons, with data acquired from ground robots, UAVs and field-mounted cameras. Corruption levels could be calibrated using actual sensor measurements and field imaging conditions. Further work should also compare different values of κ, investigate adaptive evidence aggregation, and examine whether the approach can be extended to object detection and weed segmentation. Real-time testing on edge devices and autonomous weeding platforms would be particularly important for assessing its practical deployment. Finally, combining natural-corruption robustness with adversarial robustness, uncertainty estimation and confidence calibration could provide a more complete assessment of model reliability under real agricultural conditions.

Declarations

Data Availability

The datasets generated during and/or analysed during the current study are available in the Hugging Face repository at https://huggingface.co/datasets/Rishi210904/Crop_Weed_classification and the CropAndWeed benchmark dataset is available in the GitHub repository at https://github.com/cropandweed/cropandweed-dataset.

Author Contributions

R.C. conceived the study, conducted field image acquisition, implemented the models, performed the experiments and analyses, and drafted the manuscript. A.A. supervised the research, facilitated computational resources, and critically revised the manuscript. V.P.C. supervised the research, contributed to the field study and interpretation of the results, and critically revised the manuscript. All authors read and approved the final manuscript.

Funding

This research received no specific external funding. Computational resources were provided through the institutional high-performance computing infrastructure of IISER Bhopal.

Competing Interests

The authors declare no competing financial or non-financial interests.

Ethics Approval and Plant Research Statement

The study involved non-destructive photographic imaging of cultivated crops and naturally occurring weeds from experimental plots at ICAR–Central Institute of Agricultural Engineering (ICAR–CIAE), Bhopal. Dr. V. P. Chaudhary, Head, Agricultural Mechanization Division (AMD), ICAR–CIAE, Bhopal, and a co-author of this study, authorized the collection of crop and weed images from the experimental plots of the Division. No plants, seeds, tissues, or other plant parts were collected, removed, preserved, or experimentally manipulated. Crop and weed species were identified during field data collection and verified by Dr. V. P. Chaudhary based on established morphological characteristics and field observations. The study complied with applicable institutional guidelines for research activities. As no plant specimens or plant parts were collected, no plant-collection permit or licence was required.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Clinical Trial Number

Not applicable.

References

1. Assessing the capability of YOLO- and transformer-based object detectors for real-time weed detection. (2025) *Precision Agriculture*. https://doi.org/10.48550/arXiv.2501.17387
2. Chaudhary R, Agarwal A, Chaudhary VP. (2026) Crop–weed classification using deep learning: A comparative study of CNNs, vision transformers, and interpretable models. *Scientific Reports* (2026). https://doi.org/10.1038/s41598-026-56488-4
3. Chaudhary R, Agarwal A, Chaudhary VP. The Smart Scythe: AI-powered weed detection in agriculture. Sci Rep. 2026. 10.1038/s41598-026-59178-3.
4. Chen J, Lu Y, Yu Q, et al. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint. 2021. https://doi.org/10.48550/arXiv.2102.04306. arXiv:2102.04306.
5. Dosovitskiy A, Beyer L, Kolesnikov A et al. (2021) An image is worth 16×16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*, Virtual. https://doi.org/10.48550/arXiv.2010.11929
6. Ferentinos KP. Deep learning models for plant disease detection and diagnosis. Comput Electron Agric. 2018;145:311–8. https://doi.org/10.1016/j.compag.2018.01.009.
7. Guo C, Pleiss G, Sun Y, Weinberger KQ. (2017) On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine* Learning *(ICML)*, Sydney, pp 1321–1330. https://doi.org/10.48550/arXiv.1706.04599
8. 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 (CVPR)*, Las Vegas, pp 770–778. https://doi.org/10.48550/arXiv.1512.03385
9. Hendrycks D, Dietterich T. (2019) Benchmarking neural network robustness to common corruptions and perturbations. In: *International Conference on Learning Representations (ICLR)*, New Orleans. https://doi.org/10.48550/arXiv.1903.12261
10. Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: A survey. Comput Electron Agric. 2018;147:70–90. https://doi.org/10.1016/j.compag.2018.02.016.
11. Li O, Liu H, Chen C, Rudin C. (2018) Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 32. https://doi.org/10.48550/arXiv.1710.04806
12. Ma C, Donnelly J, Liu W, Vosoughi S, Rudin C, Chen C. Interpretable image classification with adaptive prototype-based vision transformers. Adv Neural Inf Process Syst. 2024;37:41447–93. https://doi.org/10.48550/arXiv.2410.20722.
13. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. (2018) Towards deep learning models resistant to adversarial attacks. In: *International Conference on Learning Representations (ICLR)*, Vancouver. https://doi.org/10.48550/arXiv.1706.06083
14. Nauta M, van Bree R, Seifert C. (2023) PIP-Net: Patch-based interpretable prediction network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2744–2753.
15. Pound MP, Atkinson JA, Townsend AJ, et al. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. GigaScience. 2017;6(10):1–10. https://doi.org/10.1093/gigascience/gix083.
16. Sa I, Ge Z, Dayoub F, Upcroft B, Perez T, McCool C. DeepWeeds: A multiclass weed species image dataset for deep learning. Sci Rep. 2019;9:2058. https://doi.org/10.1038/s41598-018-38343-3.
17. Saralajew S, Rana A, Villmann T, Shaker A. (2025) A robust prototype-based network with interpretable RBF classifier foundations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Philadelphia. arXiv:2412.15499. https://doi.org/10.48550/arXiv.2412.15499
18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, pp 618–626. https://doi.org/10.1007/s11263-019-01228-7
19. Zhao Y, Wang X. (2026) Deep learning–based approaches for weed detection in crops: A review. *Frontiers in Plant Science*, in press. https://doi.org/10.3389/fpls.2025.1746406

Figures

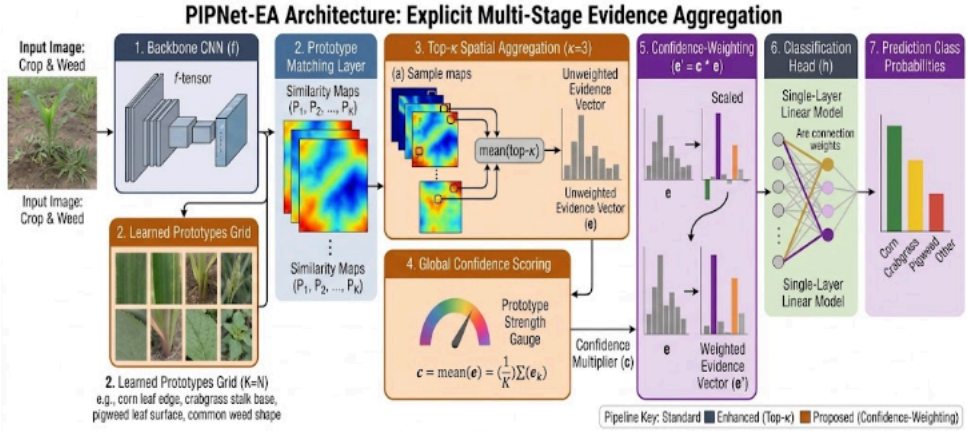


Figure 1

PIPNet-EA Architectural Pipeline and Innovations**.** Diagram illustrating (A) Feature Visualization: Input crop/weed images and extracted feature heatmaps. (B) Prototype Learning: Grid of learned crop and weed texture prototypes and similarity maps. (C) Evidence Weighing Innovations: Top-κ spatial aggregation (k=3), global confidence scoring (prototype strength gauge), and confidence-based evidence rescaling. (D) Sparse Classification: Single-layer linear model for interpretable, sparse prediction class probabilities.

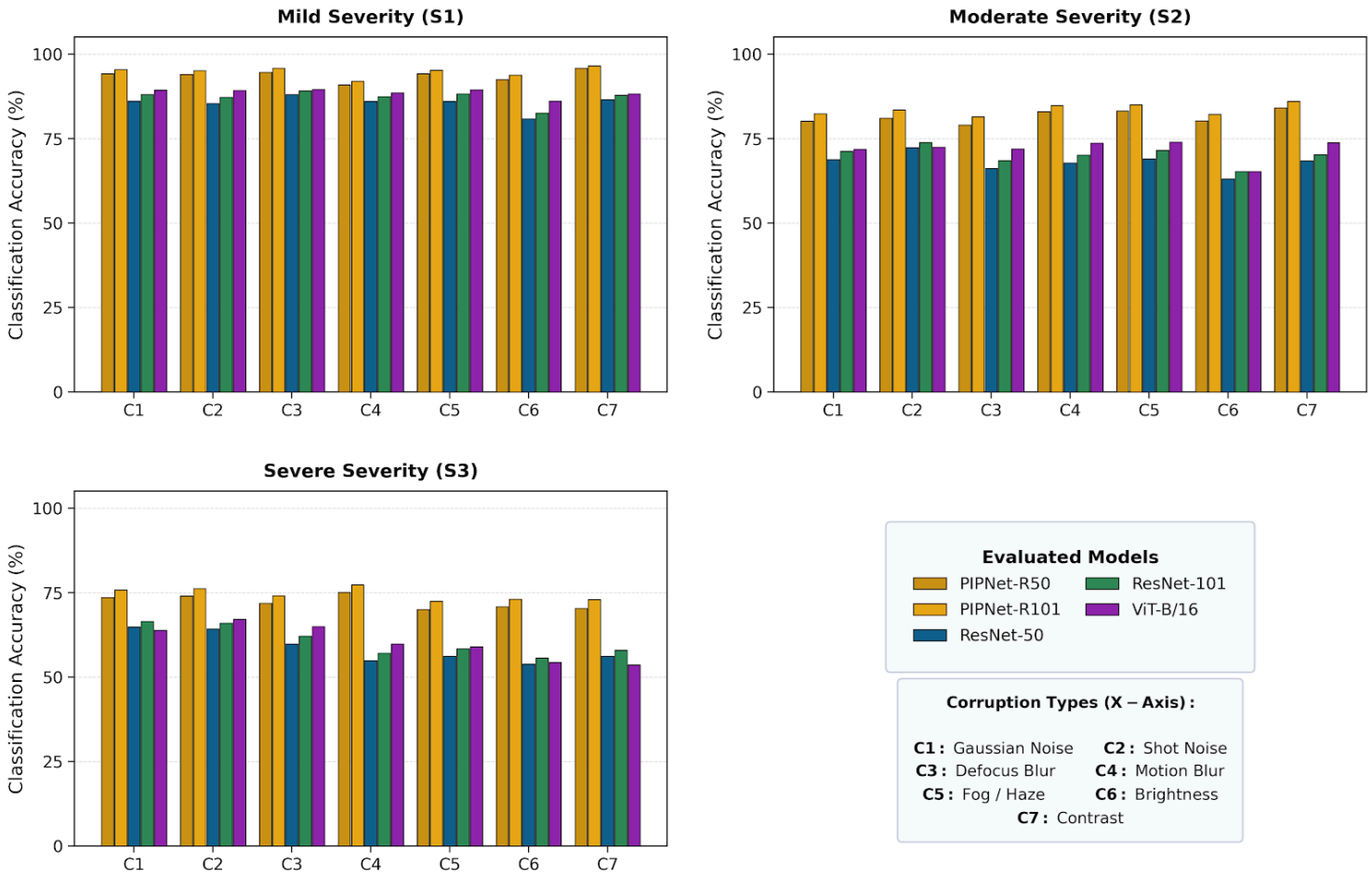


Figure 2

Classification accuracy on D1 for five models under seven corruption types and three severity levels. Left: mild (S1); centre: moderate (S2); right: severe (S3). PIPNet-R50 shows the highest accuracy across the majority of corruption–severity combinations.

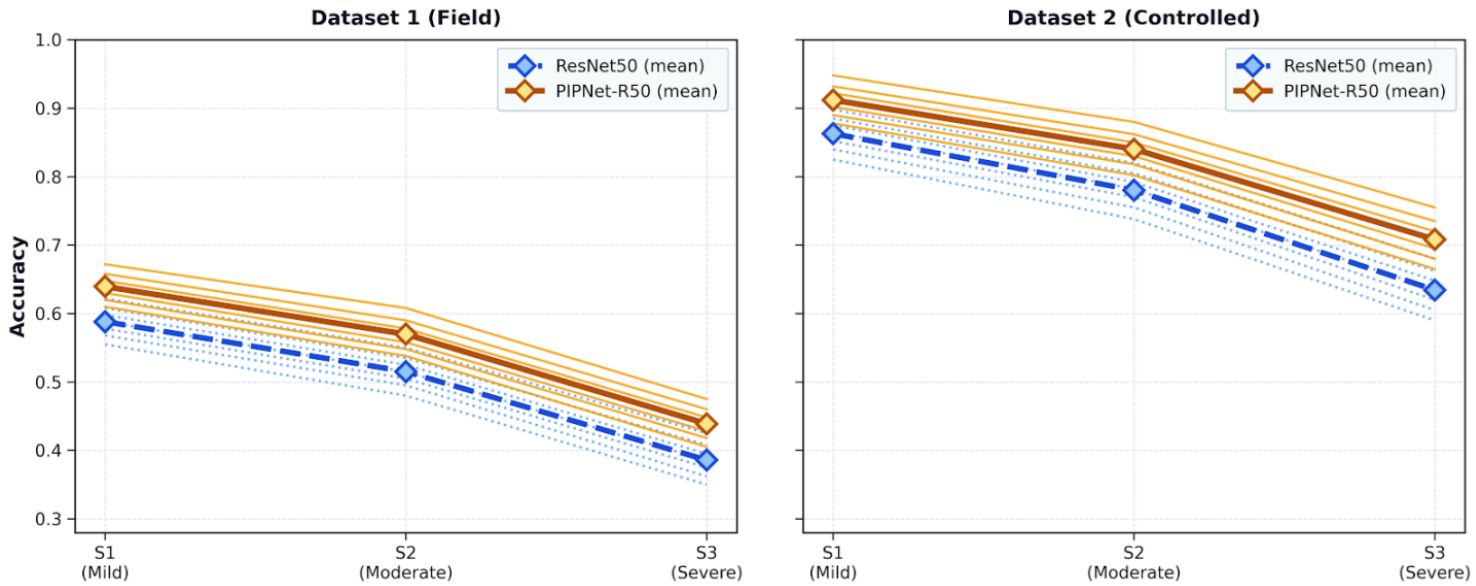


Figure 3

Accuracy degradation trajectories of PIPNet-R50 and ResNet50 from mild (S1) to severe (S3) corruption on D1 and D2. Thin lines represent individual corruption types, and thick lines represent the mean across all seven corruptions. PIPNet-R50 maintains higher accuracy than ResNet50 across the severity range on both datasets.

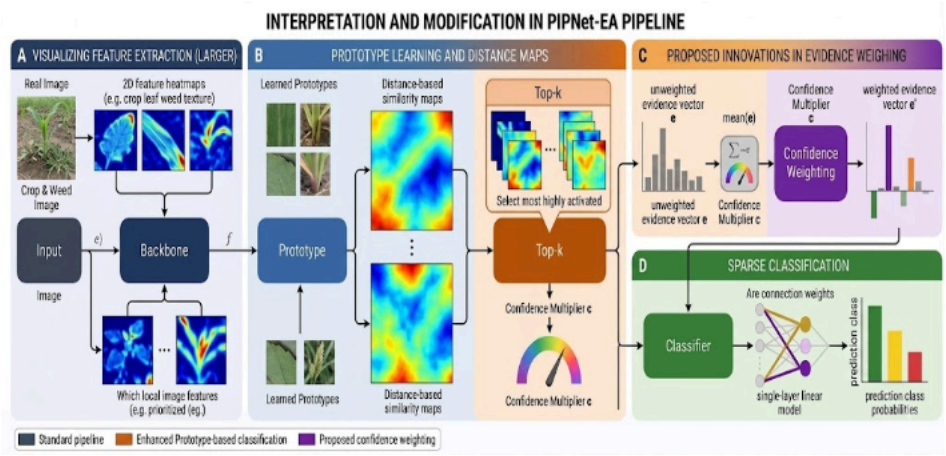


Figure 4

PIPNet-EA classification pipeline. The top-κ spatial evidence aggregation and confidence-weighting steps are the two additions to the original PIPNet framework.

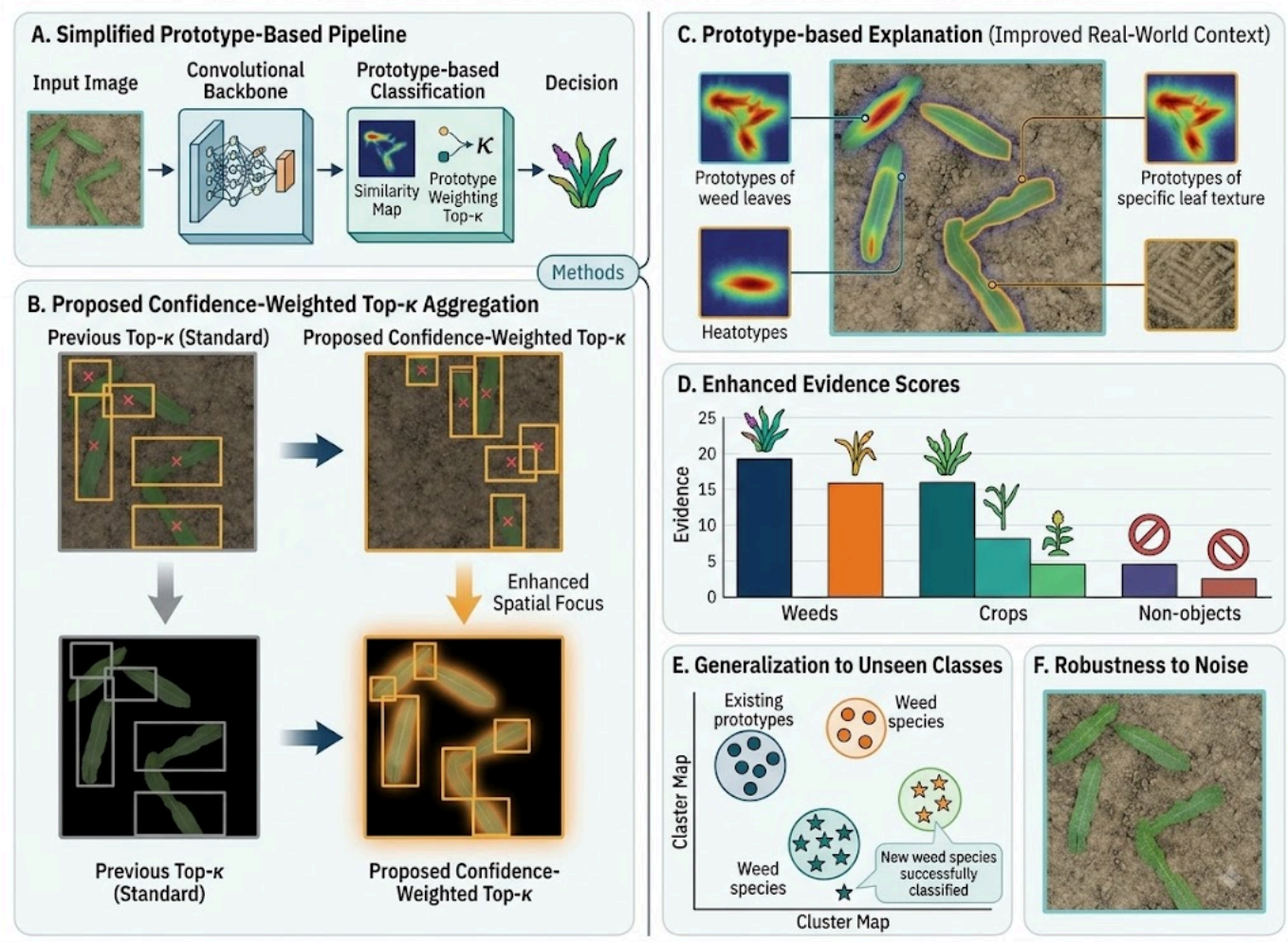


Figure 5

Prototype-based explanation for a weed-classified sample. The activation maps highlight image regions associated with weed morphology, followed by the top-κ evidence regions used for classification and the original input image. The class evidence scores indicate stronger evidence for weed (26.4) than crop (22.1).