

MSANet: Multi-Scale Attention Network for Wheat Disease Classification

Bekzat Ongdassynov
Nazarbayev University

bekzat.ongdassynov@nu.edu.kz

Rakhat Myrzakhan
Nazarbayev University

rakhat.myrzakhan@nu.edu.kz

Tairzhan Kassenov
Nazarbayev University

tairzhan.kassenov@nu.edu.kz

Denis Ten
Nazarbayev University

denis.ten@nu.edu.kz

Abstract

Early identification of wheat foliar diseases is vital for reducing yield loss, yet manual inspection is slow and inconsistent. We address automated disease recognition from RGB leaf images using a 15-class classification setup. Our model, MSANet, integrates a lightweight CNN backbone with multi-scale feature fusion and CoordAttention to capture fine-grained texture patterns and global spatial dependencies. We evaluate the method on a wheat disease dataset containing 14K training images and balanced test splits. Using accuracy and macro-F1 as metrics, MSANet reaches 90% top-1 accuracy and a macro-F1 of 0.88, outperforming our baselines. Visualization shows that the model consistently attends to lesion regions, indicating effective feature extraction. The results demonstrate that multi-scale attention improves robustness for practical wheat disease classification.

1. Introduction

Wheat is one of the most widely cultivated cereal crops, and foliar diseases such as yellow and brown rust, and leaf blotch cause significant annual yield losses. Early stage identification is essential for targeted treatment, but manual field scouting is slow, labor-intensive, and prone to human error. Automated computer vision systems offer a scalable alternative, however, wheat diseases often exhibit subtle visual differences, high intra-class variability, and varying illumination and background conditions. These factors make reliable classification challenging for standard convolutional models.

In this project, we investigate supervised wheat disease recognition using RGB leaf images and formulate the task as a 15-class classification problem. Our goal is to develop an architecture that can capture both fine-grained texture cues and higher-level semantic patterns that distinguish vi-

sually similar diseases. To this end, we propose MSANet, a multi-scale attention network that combines a lightweight convolutional backbone with feature fusion and CoordAttention to model long-range spatial dependencies.

We evaluate our method on publicly available Wheat Plant Diseases dataset [9] and compare its performance with the top models reported on the Kaggle leaderboard for this task, namely ResNet50V2 and ConvNeXt. Although these external benchmarks are not retrained locally, they provide a consistent reference point for assessing the effectiveness of our architecture. The results show that multiscale feature aggregation combined with directional CoordAttention yields competitive accuracy while offering clearer spatial focus on lesion regions. Beyond quantitative performance, qualitative analysis using Grad-CAM and attention visualizations demonstrate that MSANet consistently attends to disease-relevant areas rather than background noise, highlighting its practical suitability for real world agricultural monitoring.

2. Literature Review

Plant disease classification from leaf images has been extensively studied using convolutional neural networks (CNNs). Early deep-learning approaches demonstrated that standard CNN backbones such as AlexNet, VGG, and ResNet outperform traditional handcrafted descriptors for crop disease identification [6, 8]. However, these architectures often struggle when disease symptoms are subtle or when classes differ primarily through fine-grained lesion textures, which are commonly observed in wheat pathology.

Transfer learning has become a dominant strategy for agricultural vision tasks. Models such as ResNet50, DenseNet121, and InceptionV3 have achieved strong results on leaf-image datasets by leveraging ImageNet-pretrained features and adapting them to plant specific domains [2, 1]. While effective, these architectures typically rely on deep semantic representations and may underutilize

small scale lesion cues important for distinguishing visually similar diseases.

Attention mechanisms have recently gained prominence due to their ability to enhance interpretability and highlight disease-relevant regions. Channel and spatial-attention modules, such as SENet [4], CBAM [10], and CoordAttention [3], have been shown to refine feature selection by emphasizing lesion areas, while suppressing background noise. Several agricultural studies report improved robustness under varying illumination and cluttered field conditions when such attention blocks are applied.

Another important direction involves multi-scale feature extraction. Architectures inspired by Feature Pyramid Networks (FPN) [5] or Inception-style parallel convolutions [7] capture both shallow texture patterns and deeper semantic structures, improving performance on fine-grained plant disease datasets. However, many existing multi-scale approaches rely on heavyweight backbones or do not explicitly model long-range positional dependencies.

Our work differs by combining multi-scale fusion with direction-aware CoordAttention integrated directly into a lightweight backbone. Instead of relying solely on deep semantic features or isolated attention modules, MSANet aggregates information across multiple receptive fields and encodes positional context along height and width. This design improves fine-grained discrimination, reduces confusion between visually similar diseases (such as wheat rust types), and remains computationally efficient for real-world agricultural deployment.

3. Dataset

We use the Wheat Plant Diseases dataset from Kaggle [9], which contains high-resolution RGB images of wheat leaves affected by various diseases.

3.1. Overview

The dataset contains approximately 14155 images. Each image corresponds to a single wheat plant or leaf and is labeled with one of several disease or pest classes. The classes include fungal diseases, rusts (leaf rust, stem rust, stripe and yellow rust), blotches, smuts, pests (aphid, mite, stem fly), and other disease types.

3.2. Class distribution and preprocessing

The dataset is imbalanced across classes, with some diseases appearing far more frequently than others. To prepare the data for training and evaluation, we apply the following steps:

- Resize all images to a uniform resolution suitable for the model (e.g., 224×224).
- Perform a stratified split to preserve class distribution. In our experiments, we used roughly 13,104 training

images, 300 validation images, and 750 test images (50 per class).

- Apply standard data augmentation (random horizontal flip, random rotation, color jitter) to improve robustness to background and lighting variation.

3.3. Rationale and constraints

This dataset is appropriate for our task because:

- It contains a wide variety of wheat diseases and pests, enabling both coarse and fine-grained classification.
- The images are collected under realistic field conditions, making the dataset suitable for practical deployment scenarios.
- As a public dataset, it provides a reproducible benchmark for comparing approaches.

However, the dataset also presents some challenges:

- Class imbalance requires careful handling through augmentation or loss weighting.
- Variations in image resolution may cause loss of fine lesion details when downsampled.
- Only image-level labels are provided, which limits the task to classification rather than lesion localization.

4. Methodology

Our approach focuses on fine-grained recognition of wheat foliar diseases using an architecture that combines a strong pretrained backbone with explicit multi-scale feature extraction and spatially-aware attention. The system consists of four main components: preprocessing, the EfficientNetV2-L backbone, multi-scale feature tapping and fusion, and per-scale CoordAttention followed by a classification head.

4.1. Input preprocessing

All images are resized to 224×224 and normalized using dataset channel statistics. During training, we apply a set of regularization techniques: random horizontal flip, color jitter, random erasing, and mixup ($\alpha = 0.2$). These augmentations aim to improve robustness to lighting variation, leaf orientation, and background clutter typical in real farm conditions.

4.2. Backbone: EfficientNetV2-L

We use EfficientNetV2-L as the feature extractor due to its strong performance on natural images and its ability to

encode both local textures and global structures. The backbone processes the input through hierarchical stages, producing multi-resolution activation maps. Early layers contain fine color and lesion details, while deeper layers encode semantic patterns such as stripe-like rust structures.

Initially, the backbone is frozen. We follow a *progressive unfreezing* schedule: deeper layers remain frozen while the classifier trains, and higher-level blocks are gradually unfrozen to stabilize adaptation.

4.3. Multi-scale feature extraction

To capture fine-grained lesion cues, we tap feature maps from three distinct depth levels of EfficientNetV2-L:

- shallow: high-resolution texture patterns (e.g., small spots, speckles),
- mid-level: mixed features combining texture and shape,
- deep: low-resolution structural features capturing disease spread patterns.

Each tapped feature map is projected to a uniform channel dimension via 1×1 convolutions and spatially aligned to the same resolution through upsampling or downsampling.

4.4. CoordAttention refinement (per scale)

For each scale, we apply CoordAttention, which separately encodes positional information along height and width. This emphasizes disease-relevant spatial regions and suppresses distractors such as stems or soil. The attention mechanism helps the model attend to lesion textures that may be subtle or occupy small regions.

4.5. Multi-scale fusion

The refined feature maps from all scales are concatenated and passed through a 3×3 convolutional fusion layer. This produces a unified representation that combines fine detail with higher-level semantics, enabling robust classification across diseases with similar appearance.

4.6. Classification head

The fused feature tensor is globally averaged and passed through a two layer fully connected head ($1024 \rightarrow 512 \rightarrow 15$). Label smoothing is applied to the cross-entropy objective to prevent overconfidence. Optimization uses AdamW with cosine annealing and learning-rate warm restarts.

4.7. Hyperparameter Search

To select suitable optimization parameters for MSANet, we performed a light grid search over learning rate and weight decay while keeping all other components fixed. All trials used the same detected feature dimensions

[32, 192, 640]. Table 1 summarizes the validation accuracy obtained for each configuration. The best-performing setting was a learning rate of 3×10^{-4} with weight decay 5×10^{-5} , which we used for all final training runs.

Learning Rate	Weight Decay	Validation Accuracy
3×10^{-4}	1×10^{-4}	0.7500
3×10^{-4}	5×10^{-5}	0.7667
1×10^{-4}	1×10^{-4}	0.7233
1×10^{-4}	5×10^{-5}	0.7500
3×10^{-5}	1×10^{-4}	0.7233
3×10^{-5}	5×10^{-5}	0.7000

Table 1. Light grid search over learning rate and weight decay.

5. Evaluation Metrics

We evaluate the performance of all models using standard multi-class classification metrics. Since the dataset includes visually similar diseases and imbalanced classes, we report both overall and class-balanced measures.

5.1. Accuracy

Top-1 accuracy measures the fraction of correctly predicted labels:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\hat{y}_i = y_i),$$

where $\hat{y}_i$ and y_i are the predicted and ground-truth labels for sample i .

5.2. Precision, recall, and F1-score

For each class c , we compute:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c},$$

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

Because individual classes differ in frequency, we report the macro-averaged score:

$$\text{MacroF1} = \frac{1}{C} \sum_{c=1}^C \text{F1}_c,$$

with $C = 15$ classes.

Macro-F1 treats each class equally and reflects performance on underrepresented diseases, making it more informative than accuracy alone.

5.3. Confusion matrix

We additionally use a confusion matrix to analyze common misclassifications. It highlights visually similar disease pairs (e.g., leaf rust against stem rust), helping identify where the model struggles.

5.4. Qualitative evaluation

To inspect the model’s decision process, we use Grad-CAM visualizations. These heatmaps reveal whether the classifier attends to lesion areas rather than irrelevant background. While the metric is not quantitative, they provide important evidence of correct spatial reasoning and failure patterns.

6. Results and Analysis

6.1. Training Dynamics

The training dynamics of MSANet are illustrated in Figures 1 and 2. The model converges smoothly with a consistent increase in accuracy and a steady reduction in training loss. Validation loss fluctuates due to heavy augmentation, but validation accuracy closely tracks training accuracy, indicating limited overfitting. The cosine-annealed learning-rate schedule (Fig. 1, right) shows warm restarts that periodically restore exploration. Accuracy and loss gaps (Fig. 2) remain small and mostly positive, confirming stable generalization.

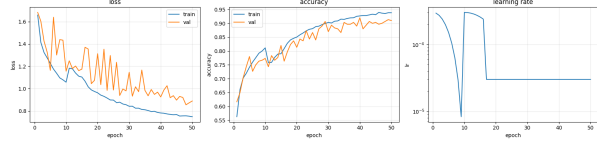


Figure 1. Training curves: loss (left), accuracy (center), and cosine-annealed learning rate with warm restarts (right).

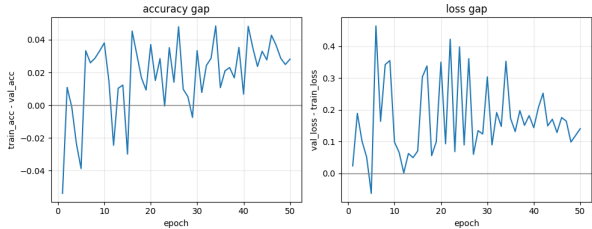


Figure 2. Generalization behavior: accuracy gap (train–val) and loss gap (val–train). Small and mostly positive gaps indicate stable training.

6.2. Comparison with external benchmarks

We compare MSANet against two top-performing architectures reported on the Kaggle leaderboard for this dataset: ResNet50V2 and ConvNeXt. We do not retrain these models locally, instead, we reference their publicly reported performance to contextualize our results. Table 2 shows the comparison between the models.

Although ConvNeXt achieves the highest accuracy, MSANet achieves a better balance between precision and recall across classes. Its multi-scale design helps focus on

Model	Accuracy	F1-score	Parameters
ResNet50V2	89.73%	0.8891	25M
ConvNeXt	90.53%	0.8802	88M
MSANet (ours)	90.00%	0.8843	126M

Table 2. Comparison of the models.

disease-specific regions, resulting in more consistent per-class F1-scores.

6.3. Per-class performance

The overall test-set performance of MSANet is summarized in Table 3. We report accuracy together with macro and weighted averages of precision, recall, and F1-score.

	Precision	Recall	F1-score	Support
Accuracy	—	—	0.9000	750
Macro avg	0.9300	0.9000	0.8800	750
Weighted avg	0.9300	0.9000	0.8800	750

Table 3. Global test-set metrics for MSANet on the held-out test set.

These global metrics show that the model maintains a strong balance between precision and recall, indicating stable performance across classes.

Figure 3 summarizes the per-class precision, recall, and F1-scores, highlighting variations in difficulty across the 15 disease categories.

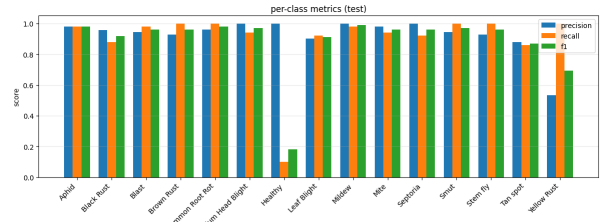


Figure 3. Per-class precision, recall, and F1-scores on the test set.

Most classes achieve precision and recall above 0.95, indicating stable recognition of visually distinctive diseases, and only *Healthy* and *Yellow Rust* classes stand out as the weakest performers.

The confusion matrix in Figure 4 illustrates the structure of misclassifications across all 15 classes.

The main difficulties arise in:

- **Healthy** — often misclassified as disease due to natural textures mimicking lesions.
- **Yellow Rust** — visually similar to healthy leaves under certain lighting conditions.

To further analyze the reliability of MSANet’s predictions, we examine the distribution of maximum predicted probabilities, shown in Figure 5.

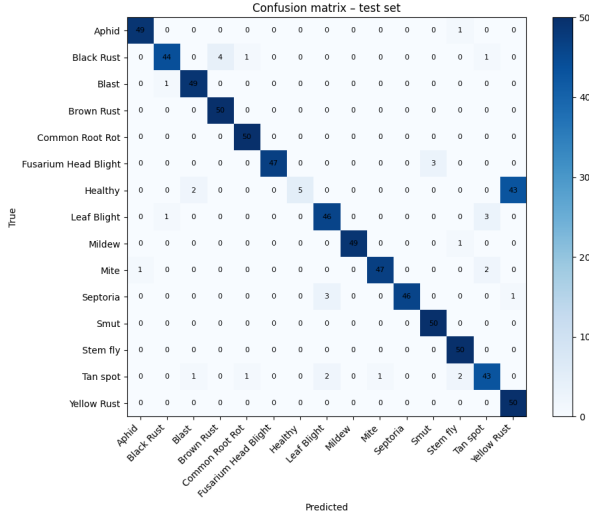


Figure 4. Confusion matrix for the 15-class test set.

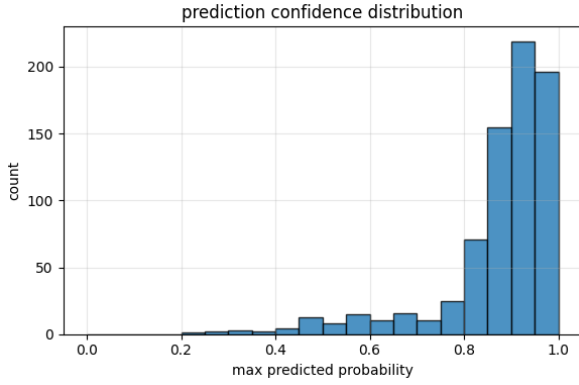


Figure 5. Distribution of the model's maximum predicted probabilities on the test set.

The histogram reveals that the majority of predictions lie in the high-confidence range (0.85–1.0), indicating that the model is typically decisive. The small fraction of low-confidence predictions aligns with classes known to be challenging - primarily *Healthy* and *Yellow Rust*.

6.4. Interpretability analysis

We visualize MSANet using Grad-CAM and multi-scale CoordAttention maps. Figure 6 illustrates, for several Aphid examples, how CoordAttention consistently highlights pest regions across scales.

The heatmaps show that:

- MSANet consistently focuses on lesion spots, pustules, and stripe patterns.
- Background elements are strongly suppressed.
- Healthy leaves trigger diffuse activation, explaining misclassifications.

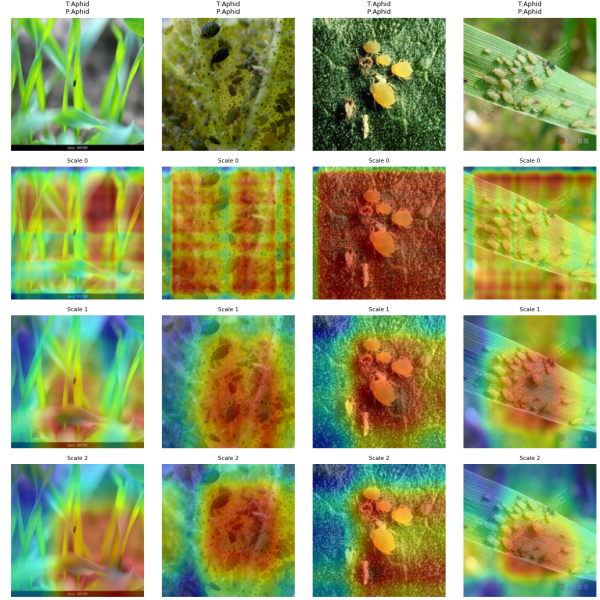


Figure 6. Multi-scale CoordAttention visualizations for the *Aphid* class. The first row shows the original images, while the following rows correspond to attention maps at three scales.

We additionally examine Grad-CAM visualizations to confirm that the final classifier focuses on lesion regions rather than background artifacts. Figure 7 shows representative examples for the *Aphid* class.

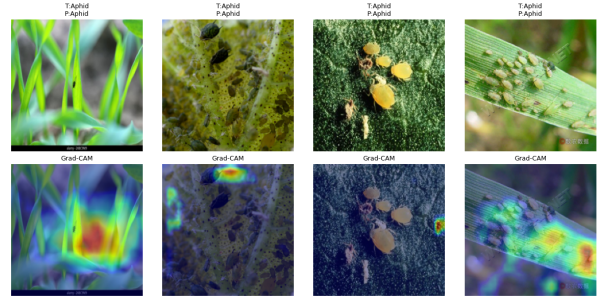


Figure 7. Grad-CAM visualizations for several *Aphid* samples. The first row presents the original images, while the second row shows the corresponding Grad-CAM heatmaps.

6.5. Discussion

MSANet provides a strong trade-off between interpretability and accuracy. Its structured attention makes it more trustworthy for agricultural deployment, even though it does not strictly outperform ConvNeXt in raw accuracy. The results indicate that wheat disease classification benefits from combining hierarchical lesion cues with explicit positional attention.

7. Error Analysis

7.1. Misclassification patterns

Analysis of the confusion matrix reveals several systematic error modes. The most prominent issue occurs in the *Healthy* class. Healthy leaves exhibit natural texture variations—light reflections, veins, and small speckles—that resemble early disease symptoms. As a result, the model frequently classifies Healthy samples as diseased, particularly as Yellow Rust. This occurs because Yellow Rust displays elongated, stripe-like patterns that visually overlap with natural highlights on Healthy leaves.

A second source of errors involves visually similar rust diseases. Yellow Rust, Brown Rust, and Black Rust share overlapping color distributions and lesion shapes. Despite the multi-scale architecture, some deep feature embeddings remain close in high-dimensional space, leading to occasional cross-confusion.

Minor classes, especially pest-related categories (e.g., aphid, mite, stem fly), also exhibit slightly lower precision. These classes have fewer samples and exhibit high intra-class variability, which limits the model’s ability to form a robust representation. To further illustrate this issue, Fig. 8 shows a representative pair of samples: a Healthy leaf that was misclassified as Yellow Rust, and an actual Yellow Rust leaf. The two images share nearly identical global color tones and surface textures. In the Healthy sample, natural vein structure and subtle reflectance patterns resemble early rust streaks, which leads the model to activate similar feature channels. This visual similarity explains why the Healthy class exhibits the lowest recall among all categories.

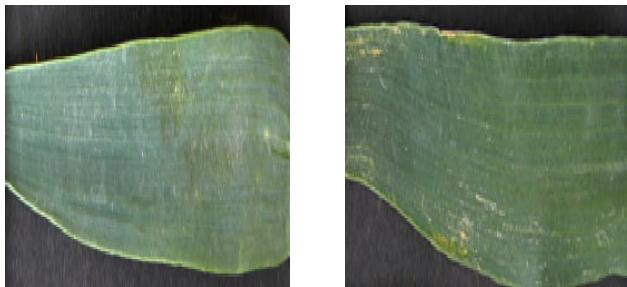


Figure 8. Visual similarity between Healthy and Yellow Rust (in order as shown) leaves contributes to systematic misclassification.

7.2. Attention behavior and spatial bias

Grad-CAM and CoordAttention visualizations provide insight into failure cases. While MSANet correctly focuses on lesion regions for most diseases, failure cases often show *diffuse or background-biased attention*. For Healthy leaves, attention maps highlight broad areas instead of specific cues, causing the classifier to interpret benign texture

as disease indicators.

For Yellow Rust the attention occasionally overemphasizes lighting patterns rather than true stripe lesions. This aligns with the confusion observed in the quantitative results and reflects limits in fine-grained discrimination between natural leaf texture and early rust symptoms.

7.3. Model bias toward caution

Another pattern visible in the confusion matrix is that the model tends to predict disease rather than predict that the wheat is Healthy. This conservative bias results in fewer false negatives at the cost of increased false positives. In agricultural scenarios, this reflects a preference for caution—misidentifying a healthy plant as diseased is less costly than failing to detect early infection, but it reduces overall accuracy on Healthy leaves.

7.4. Root causes and contributing factors

Across these error modes, several root causes can be identified:

- **Insufficient diversity in Healthy samples:** natural variability of lighting, age, and environmental artifacts is underrepresented.
- **Texture similarity between early disease and natural patterns:** especially for stripe-like diseases such as Yellow Rust.
- **Minor-class imbalance:** small pest categories provide limited signal for robust feature learning.
- **Lack of lesion-level supervision:** without segmentation masks, spatial attention must infer lesion regions indirectly, sometimes amplifying noise.

7.5. Potential improvements

Based on these observations, several directions could reduce errors:

- Increasing diversity of Healthy samples to better model natural leaf variability.
- Reducing augmentations that possibly unintentionally introduce disease-like textures.
- Exploring contrastive pre-training (Healthy against Diseased) to improve discriminative boundaries.
- Adding lesion segmentation or mask supervision to prevent diffuse attention.
- Introducing an “uncertain” class to avoid forced disease predictions in ambiguous cases.

Overall, the error patterns confirm that MSANet captures most disease-related cues effectively but struggles with classes where natural texture closely resembles pathology or where sample imbalance limits representational strength.

8. Future Work

Several directions could further improve the performance and robustness of MSANet in real agricultural settings. First, incorporating lesion-level supervision, either through manual masks or weakly-supervised segmentation, may reduce diffuse attention and guide the model toward true disease regions. This is particularly relevant for rust diseases where spatial structure plays a critical role.

Second, class imbalance remains a challenge, especially for minority pest classes. Techniques such as class-balanced sampling, Focal Loss, or synthetic sample generation could help stabilize training. Related to this, expanding the diversity of Healthy samples may reduce false positives caused by natural leaf texture.

Third, the current backbone is computationally heavy. Exploring lighter architectures, such as EfficientNetV2-S or MobileNet-based variants, would enable deployment on edge devices used in field environments. Complementary strategies such as quantization or pruning could further reduce inference cost.

Another promising direction is contrastive or self-supervised pretraining tailored to agricultural imagery. Representations learned from Healthy vs. Diseased contrast might improve separation between visually similar categories.

Finally, introducing an uncertainty-aware prediction mechanism (e.g., threshold-based abstention or an explicit “uncertain” class) could reduce forced misclassifications in borderline cases. Such mechanisms are valuable in precision agriculture, where downstream decisions depend heavily on model confidence.

Overall, addressing these aspects would improve scalability, interpretability, and reliability, making MSANet more suitable for deployment in real world crop monitoring systems.

References

- [1] J. G. A. Barbedo. Plant disease identification from individual lesions and spots using deep learning. *Biosystems Engineering*, 180:96–107, 2019.
- [2] K. P. Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018.
- [3] Q. Hou, D. Zhou, and J. Feng. Coordattention: Efficient channel attention with coordinate information. In *CVPR*, 2021.
- [4] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.
- [5] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.
- [6] S. P. Mohanty, D. P. Hughes, and M. Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016.
- [7] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. In *CVPR*, 2015.
- [8] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun. A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, 161:272–279, 2019.
- [9] K. Verma. Wheat plant diseases dataset. <https://www.kaggle.com/datasets/kushagra3204/wheat-plant-diseases>, 2024. Accessed: 2025-12-03.
- [10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.