Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 3 (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Advanced Deep Learning Framework for Soybean Leaf Disease Detection, Classification and Segmentation Using UAV Imagery

Abhishek Kumar Agrawal[1*], Anup Mishra[2], Mukesh Kumar Chandrakar[3], Abhishek Verma[4]

[1234]Department of Electrical & Electronics Engineering, Bhilai Institute of Technology Durg, Durg, Chhattisgarh, India.

***Corresponding author: Abhishek Kumar Agrawal**

Email ID: abhishek.agarwal@bitdurg.ac.in

Cite this paper as: Abhishek Kumar Agrawal, Anup Mishra, Mukesh Kumar Chandrakar, Abhishek Verma (2025) Advanced Deep Learning Framework for Soybean Leaf Disease Detection, Classification and Segmentation Using UAV Imagery. Journal of Neonatal Surgery, 14, (3) 374-387

ABSTRACT

Accurate and early detection of plant diseases is critical for sustainable agriculture and crop yield optimization. This study presents a unified deep learning framework for soybean disease classification, anomaly detection, and spatial localization using high-resolution UAV-based imagery. We adopt an attention-based Multi-Instance Learning (MIL) approach for image-level disease classification, enabling the model to focus on disease-relevant regions within heterogeneous field scenes using only image-level supervision. To detect both known and previously unseen disease patterns, we integrate a memory-based patch level anomaly detection mechanism that models healthy soybean appearance in feature space and identifies deviations via nearest-neighbor distances. Additionally, we employ a self-supervised contrastive segmentation pipeline to generate pixel-wise disease localization maps without requiring manual annotations. The proposed framework addresses key challenges in real-world agricultural monitoring, including label scarcity, mixed health states, and complex backgrounds. Extensive experiments demonstrate that the integration of MIL-based classification, memory-based anomaly detection, and self-supervised segmentation enables robust, scalable, and interpretable disease monitoring from UAV imagery, making the framework suitable for precision agriculture applications.

Keywords: *Precision Agriculture, AI, Plant Disease Detection, Segmentation*

1. INTRODUCTION

Plant diseases pose a persistent and significant threat to global food security, with estimated yield losses reaching up to 40% in major crops if not managed in a timely manner [1, 2]. Conventional disease monitoring practices rely on manual field scouting and expert visual inspection, which are labor-intensive, subjective, and impractical for large-scale agricultural systems. The growing demand for sustainable and precision agriculture has therefore motivated the adoption of automated disease detection systems based on computer vision and deep learning.

Early deep learning approaches for plant disease identification primarily focused on image-level classification using convolutional neural networks (CNNs), achieving high accuracy on benchmark datasets composed of isolated leaf images captured under controlled conditions [3, 4]. Subsequent studies introduced deeper architectures and attention mechanisms to improve robustness and generalization [5, 6]. Despite these advances, most existing methods implicitly assume that disease symptoms are uniformly distributed across an image. This assumption rarely holds in real-world agricultural settings, where disease symptoms often appear sparsely and coexist with healthy vegetation, soil, shadows, and weeds. As a result, global image classifiers may overlook subtle disease cues or generate overconfident predictions driven by background artifacts.

Unmanned Aerial Vehicles (UAVs) have emerged as a powerful tool for large scale crop monitoring by enabling rapid, non-invasive, and high-resolution imaging of agricultural fields [7, 8]. UAV-based imagery captures rich spatial context and disease spread patterns that are unavailable in single-leaf datasets, making it well suited for early disease detection and precision intervention. However, UAV imagery introduces additional challenges, including mixed healthy and diseased

regions within a single image, variable illumination, scale variation, occlusion, and complex backgrounds [9, 10]. These factors significantly limit the effectiveness of conventional Image level classification models.

In parallel, anomaly detection has been explored as an alternative paradigm for plant disease identification, particularly for detecting rare or previously unseen disease patterns. Reconstruction-based methods such as auto encoders and variational models learn representations of healthy plants and flag deviations as anomalies [11]. While effective in controlled environments, these approaches often produce blurred reconstructions and exhibit weak localization performance in complex agricultural scenes. More recently, memory-based and distance-based anomaly detection methods have demonstrated superior localization accuracy by operating directly in feature space rather than pixel space [12]. Nevertheless, their application to UAV-based crop disease monitoring remains relatively underexplored.

To address these challenges, we propose a unified deep learning framework that integrates image-level classification, unsupervised anomaly detection, and pixel-wise disease localization within a single pipeline. For classification, we adopt an attention based Multi-Instance Learning (MIL) formulation that treats each UAV image as a bag of local patches and learns to focus on disease-relevant regions using only image-level labels. This formulation explicitly models the heterogeneous nature of field imagery and provides inherent interpretability through learned attention weights. For anomaly detection, we employ a patch-level memory-based feature embedding approach that models healthy soybean appearance in feature space and detects deviations using nearest-neighbor distances, enabling robust detection of both known and unknown disease manifestations without requiring diseased annotations. Finally, a self-supervised contrastive segmentation module is used to localize disease-affected regions at the pixel level, guided by the fused outputs of classification and anomaly detection.
We validate the proposed framework on UAV-acquired soybean imagery, a crop of major global importance that is highly susceptible to visually similar foliar diseases and pest damage [13, 14]. The main contributions of this work are summarized as follows:

- We introduce an attention-based Multi-Instance Learning framework for UAV-based soybean disease classification that effectively handles sparse and heterogeneous disease patterns using only image-level supervision.
- We integrate a memory-based patch-level anomaly detection mechanism capable of identifying both known and previously unseen disease patterns without requiring pixel-level or disease-specific annotations.
- We combine classification confidence and anomaly evidence through a unified fusion strategy to guide self-supervised segmentation, resulting in interpretable and spatially precise disease localization under real-world field conditions.

2. Related Works

2.1 Plant Disease Classification

Automated plant disease classification has been extensively studied using deep learning, particularly with convolutional neural networks (CNNs) that learn hierarchical feature representations from raw images. Early works demonstrated strong performance using architectures such as AlexNet, VGG, and ResNet on datasets containing isolated leaf images captured under controlled laboratory conditions [3, 4]. Subsequent studies incorporated deeper networks, transfer learning, and attention mechanisms to improve robustness and accuracy [5, 15]. However, these approaches often struggle to generalize to real-world agricultural environments characterized by complex backgrounds, occlusions, and variable illumination.

Transformer-based architecture has recently been introduced to capture long range spatial dependencies and global contextual information [16]. While such models improve resilience to spatial variability, they continue to treat each image as a single homogeneous entity. This design limits their effectiveness for UAV imagery, where disease symptoms may occupy only a small fraction of the image and coexist with healthy vegetation. To address this limitation, weakly supervised learning paradigms such as Multi-Instance Learning (MIL) have gained increasing attention. MIL formulations model each image as a collection of instances and infer image-level labels by selectively attending to discriminative regions. Attention-based MIL has shown strong performance in medical imaging and remote sensing tasks, where localized abnormalities are embedded within large images, but its application to UAV-based plant disease detection remains limited.

2.2 Anomaly Detection for Plant Health Monitoring

Anomaly detection offers a complementary approach to disease identification by modeling normal plant appearance and detecting deviations indicative of stress or infection. Autoencoder-based methods and variational approaches reconstruct healthy samples and use reconstruction error as an anomaly score [11]. Although effective in controlled settings, these methods often produce blurred reconstructions and suffer from limited localization accuracy when applied to complex agricultural scenes with high background variability.
Memory-based and distance-based anomaly detection methods have recently emerged as strong alternatives. These approaches

store representative feature embeddings of normal samples and compute anomaly scores using nearest-neighbor distances in feature space [12]. Patch-level memory methods, in particular, provide sharp localization of anomalous regions and are well suited for high-resolution imagery where disease symptoms are spatially localized. Despite their success in industrial inspection and medical anomaly detection, their adoption in UAV-based crop disease monitoring has been relatively sparse.

2.3 Segmentation without Dense Annotations

Pixel-wise localization of plant diseases is typically addressed using fully supervised segmentation networks such as U-Net and DeepLab, which require dense, pixel-level annotations [17]. The high cost and subjectivity of annotation limit their scalability in large agricultural datasets. To reduce annotation requirements, recent research has explored weakly supervised and self-supervised segmentation methods based on contrastive learning and clustering [18–20]. These methods learn semantically meaningful feature representations without explicit labels and generate segmentation maps via clustering in embedding space. However, their effectiveness often depends on strong upstream signals to associate clusters with disease semantics.
Our work bridges these research directions by integrating attention-based Multi- Instance Learning for classification, memory-based patch-level anomaly detection, and self-supervised segmentation within a unified framework. This integration enables robust disease identification, detection of unknown anomalies, and interpretable spatial localization under realistic UAV imaging conditions.

2.3 Segmentation without Dense Annotations

Pixel-wise localization of plant diseases is typically addressed using fully supervised segmentation networks such as U-Net and DeepLab, which require dense, pixel-level annotations [17]. The high cost and subjectivity of annotation limit their scalability in large agricultural datasets. To reduce annotation requirements, recent research has explored weakly supervised and self-supervised segmentation methods based on contrastive learning and clustering [18–20]. These methods learn semantically meaningful feature representations without explicit labels and generate segmentation maps via clustering in embedding space. However, their effectiveness often depends on strong upstream signals to associate clusters with disease semantics.
Our work bridges these research directions by integrating attention-based Multi- Instance Learning for classification, memory-based patch-level anomaly detection, and self-supervised segmentation within a unified framework. This integration enables robust disease identification, detection of unknown anomalies, and interpretable spatial localization under realistic UAV imaging conditions.

3 Methodology

3.1 Overview of the Proposed Framework

We propose a unified deep learning framework for robust soybean disease detection using high-resolution UAV-acquired RGB imagery. The framework integrates three complementary components: an attention-based Multi-Instance Learning (MIL) classifier for image-level disease recognition, a memory-based patch-level anomaly detection module for identifying abnormal regions, and a self-supervised segmentation module for fine-grained pixel-wise disease localization, as illustrated in Figure 1.

In the classification stage, each UAV image is decomposed into local patches and treated as a bag of instances. An attention-based MIL model aggregates patch-level features to produce image-level disease predictions while highlighting disease-relevant regions. In parallel, the anomaly detection module models healthy soybean appearance in feature space using a patch-level memory bank and identifies deviations via nearest neighbor distances, enabling detection of both known and unknown disease patterns without requiring diseased annotations.

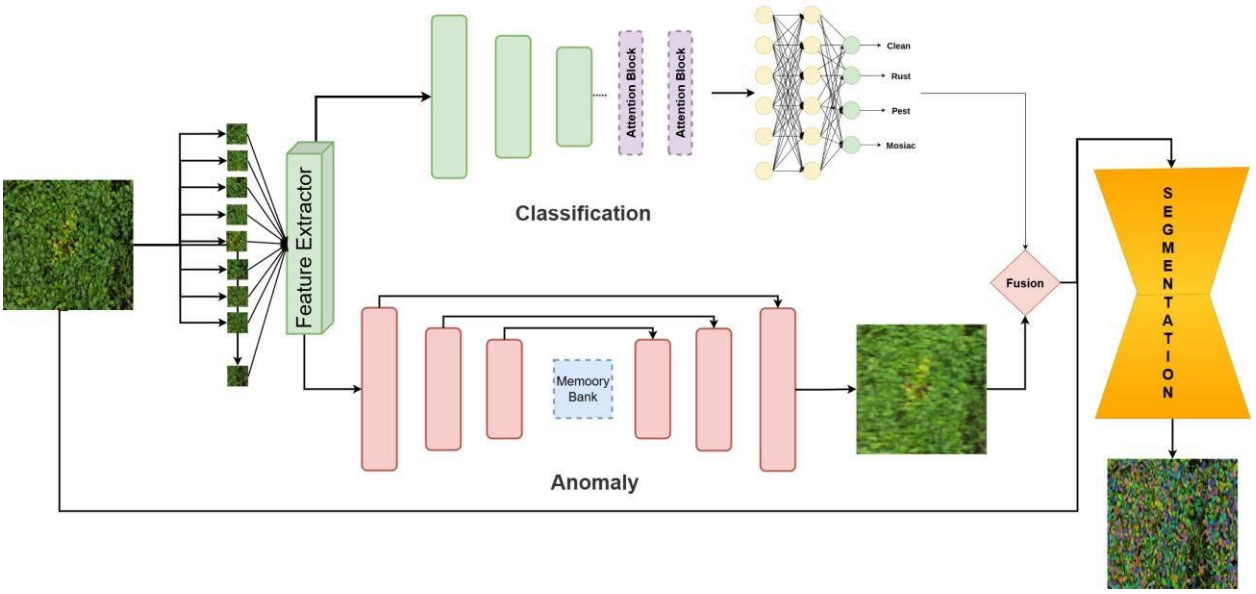

Fig. 1 Overview of the proposed framework integrating attention-based Multi-Instance Learning for classification, memory-based patch-level anomaly detection, and self-supervised segmentation.

The outputs of the MIL classifier and anomaly detector are fused into a unified confidence score that reflects both semantic class evidence and visual abnormality. This score serves as a gating signal for the subsequent self-supervised segmentation module, which leverages contrastive representation learning and clustering to generate pixel wise disease localization maps. The overall framework achieves robust classification, effective anomaly detection, and interpretable spatial localization under realistic UAV imaging conditions.

3.2 Problem Formulation

Let $I = \{I_1, I_2, \ldots \ldots, I_N\}$ denote a dataset of N UAV-acquired RGB images of soybean fields, each image $I_i \in \mathbb{R}^{H\times W\times 3}$ representing a high-resolution aerial view. Our objective is to develop a robust framework that jointly performs (i) disease classification at the image level, (ii) anomaly detection to flag unknown or novel disease phenotypes, and (iii) pixel-wise localization of symptomatic regions. Formally, the goal is to learn a mapping:

$$F : I \rightarrow (\hat{y}, \hat{a}, \hat{S})$$

where:

- $\hat{y} \in C$ is the predicted disease class from a predefined set C (including healthy),
- $\hat{a} \in [0,1]$ is an anomaly score estimating deviation from healthy patterns,
- $\hat{S} \in \{0,1\}^{H\times W}$ is a segmentation mask highlighting disease-affected pixels.

This composite prediction enables both coarse-grained (classification) and fine grained (localization) disease monitoring from UAV data without relying entirely on pixel-level annotations. Unlike purely supervised approaches, our framework integrates supervised, unsupervised, and self-supervised learning paradigms to address field-level variability, dataset limitations, and annotation costs, in line with prior works on hybrid plant disease detection [16, 18, 21].

3.3 Image-Level Classification via Attention-Based Multi-Instance Learning

Unlike conventional image classification models that assume disease symptoms are uniformly distributed across the image, UAV-acquired agricultural imagery often contains heterogeneous regions where diseased and healthy plants coexist. To address this challenge, we formulate disease classification as a Multi-Instance Learning (MIL) problem, where each UAV image is treated as a bag of instances rather than a single holistic sample.

Given an input image $I_i \in \mathbb{R}^{H\times W\times 3}$, we divide it into a set of K overlapping patches $\{x_1, x_2, \ldots \ldots, x_K\}$ using a sliding window strategy. Each patch represents an instance within a bag corresponding to the full image. A convolutional backbone network $f_\theta(.)$ (EfficientNet-B0 in our implementation) extracts a feature embedding $h_k = f_\theta(x_k)$ for each patch.

To aggregate instance-level features into an image-level prediction, we employ an attention-based pooling mechanism. The attention weight α_k for each instance is computed as:

$$\alpha_k = \frac{\exp\left(w^T \tanh\left(Vh_k\right)\right)}{\sum_{j=1}^{K} \exp\left(w^T \tanh\left(Vh_j\right)\right)} \quad (1)$$

where V and w are learnable parameters. The final bag-level representation is obtained as:

$$z = \sum_{k=1}^{K} \alpha_k h_k \quad (2)$$

The aggregated feature z is passed through a fully connected layer to produce class probabilities $\hat{y} \in \mathbb{R}^{|c|}$.This attention mechanism enables the model to focus on disease-relevant regions while suppressing background noise such as soil, weeds, or shadows

The MIL-based classifier is trained using standard cross-entropy loss with image-level labels only, making it particularly suitable for UAV-based disease monitoring where pixel-level annotations are unavailable. Moreover, the learned attention weights provide inherent interpretability by highlighting patches that contribute most strongly to the disease prediction.

3.4 Patch-Level Anomaly Detection via Memory-Based Feature Embedding

To enable robust detection of known and unknown disease patterns without requiring labeled anomalous samples, we adopt a memory-based anomaly detection approach inspired by Patch Core. Instead of reconstructing images as in autoencoder-based methods, this technique models the distribution of healthy plant appearances directly in feature space.

During training, only healthy soybean images are used. Each image is divided into local patches, and deep feature embeddings are extracted using a pretrained convolutional backbone (ResNet-50). Let $f_{i,j} \in \mathbb{R}^d$ denote the feature embedding corresponding to the patch at spatial location (i, j). All embeddings extracted from healthy images are stored in a memory bank M after dimensionality reduction using random projection to reduce redundancy.

At inference time, given a test image I, patch-level features $f_{i,j}$ are extracted and compared against the memory bank using nearest-neighbor search. The anomaly score at each spatial location is defined as:

$$A_{ij} = \min_{m \in \mathrm{M}} ||f_{i,j} - m||_2 \quad (3)$$

High anomaly scores indicate patches whose visual patterns deviate significantly from healthy soybean foliage, corresponding to potential disease symptoms or pest damage. A global image-level anomaly score â is obtained by aggregating the top percentile of patch-level anomaly scores:

$$\hat{a} = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} A_{i,j} \quad (4)$$

where Ω denotes the set of patches with anomaly scores above the 95th percentile.
This memory-based approach offers several advantages over reconstruction-based auto encoders: it avoids blurry reconstructions, provides sharper localization of anomalies, and remains robust under background clutter and varying illumination− conditions commonly encountered in UAV imagery. Importantly, it enables detection of previously unseen disease patterns without requiring any retraining or labeled anomalous data.

3.5 Fusion Strategy and Confidence Aggregation

To robustly estimate disease presence and severity under heterogeneous field conditions, we design a confidence-aware fusion strategy that integrates the complementary outputs of the attention-based Multi-Instance Learning (MIL) classifier and the patch level memory-based anomaly detector. While the MIL classifier provides high-level semantic predictions using image-level supervision, the anomaly detector captures fine grained visual irregularities indicative of known and unknown disease manifestations. Their fusion enables reliable inference in both in-distribution and out-of-distribution scenarios.
Let $\hat{y} \in \mathbb{R}^{|c|}$ denote the softmax-normalized class probability vector produced by the MIL classifier and let $\hat{a} \in [0, 1]$represent the normalized global anomaly score derived from patch-level nearest-neighbor distances. We compute a unified confidence score $\hat{c} \in \mathbb{R}^{|c|}$ as a weighted combination of these two signals:

$$\hat{c} = \alpha \,.\, \hat{y} + (1 - \alpha) \,.\, \hat{a} \ , \tag{5}$$

where $\alpha \in [0,1]$ is a tunable fusion coefficient controlling the relative influence of semantic classification and anomaly evidence. In this formulation, the anomaly score $\hat{}$a is broadcast across the class dimension, acting as a disease presence prior that modulates the classifier confidence. Higher anomaly values amplify the likelihood of disease-related classes, while suppressing overconfident predictions on visually ambiguous or previously unseen patterns.

The fused confidence score $\hat{}$c serves three key purposes within the proposed framework. First, it enables robust decision-making by mitigating failure cases where the classifier exhibits high confidence despite abnormal visual cues. Second, it facilitates threshold-based rejection of uncertain predictions, allowing samples with low confidence to be flagged for manual inspection or deferred diagnosis. Third, it acts as a gating signal for the subsequent segmentation module, ensuring that computationally intensive pixel-wise localization is activated only when there is sufficient evidence of disease presence.
This fusion mechanism effectively bridges discriminative and distance-based reasoning, combining global semantic understanding with local deviation awareness. As a result, the proposed framework achieves improved reliability, interpretability, and generalization under real-world UAV imaging conditions characterized by background clutter, mixed health states, and distributional shifts.

3.6 Pixel-wise Segmentation via Self-Supervised Contrastive Learning

Following disease detection by the fused classification-anomaly scoring module, we activate a segmentation stage to spatially localize disease symptoms at the pixel level within UAV-captured RGB images. This segmentation module is conditionally invoked only when the aggregated confidence score ĉ surpasses a predefined threshold, thereby reducing unnecessary computation on healthy samples and improving scalability. We adopt a contrastive self-supervised learning approach to pretrain a ResNet-50 encoder f_θ using the SimCLR framework [22], and alternatively evaluate MoCo [23] and BYOL [24] for robustness. Each UAV image is augmented into two distinct views through random cropping, color jittering, Gaussian blur, and horizontal flipping. The encoder learns to maximize agreement between augmented views using the normalized temperature-scaled cross entropy (NT-Xent) loss:

$$\mathcal{L}_{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(f_\theta(x_i), f_\theta(x_j))/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(f_\theta(x_i), f_\theta(x_k))/\tau)}, \tag{6}$$

where sim(·,·) denotes cosine similarity, τ is the temperature parameter, and x_i and x_j are positive pairs derived from the same image. This learning enforces semantic consistency in visual representations without requiring any labels. Once the encoder is pretrained, we extract the spatial feature map $Z = f_\theta(I) \in \mathbb{R}^{H' \times w' \times d}$ for each image. To derive a pixel-wise segmentation, we perform K-means clustering over the H′ × W′ pixel embeddings, with K = 3 representing healthy crops, diseased regions, and background. Cluster assignments are then projected back to the image grid, forming a coarse segmentation mask $M \in \{0,1,2\}^{H' \times W'}$.

To assign semantic meaning to each cluster, we correlate cluster centroids with the anomaly heatmap A∈ $\mathbb{R}^{H' \times w'}$ generated by the patch-level memory-based anomaly detection module. The cluster with the highest average anomaly score is labeled as the

diseased region, while others are mapped to healthy vegetation and background, respectively. The initial segmentation mask is further refined using Dense Conditional Random Fields (CRFs) with Gaussian bilateral potentials for appearance and position to preserve edge alignment and semantic coherence. Additionally, morphological operations such as opening and closing are applied to remove noise and smooth the boundaries of the diseased regions. The final disease mask $M_{disease}$ captures fine-grained lesion boundaries, suitable for downstream spatial diagnostics and precision intervention. This segmentation module is modularly integrated as the final step in our detection framework. It operates only when $\hat{c}$ indicates confident disease presence, leveraging prior classification output $\hat{y}$ and anomaly score $\hat{a}$ as gating mechanisms. This design ensures both high-resolution localization and computational efficiency, all while maintaining label efficiency by circumventing the need for pixel-wise supervision.

4 Dataset Preparation

The dataset utilized in this study is sourced from a publicly available repository on Mendeley Data [25], specifically curated for research in plant pathology and precision agriculture. It comprises high-resolution RGB images of soybean leaves captured under diverse natural lighting and environmental conditions, reflecting real-world field variability. Each image is labeled based on the visible presence of biotic stressors such as fungal infections or pest-induced physical damage. As shown in Figure 2, soybean leaves exhibit a variety of diseases, and it is particularly challenging to differentiate between rust and mosaic due to their visually similar symptoms. Moreover, traces of pest attacks are difficult to detect in UAV imagery, as the resulting holes appear very small from aerial views. Despite these adverse conditions, our proposed technique demonstrates strong performance and reliably identifies the affected regions.

Fig. 2 Representative UAV-captured soybean leaf samples illustrating various foliar diseases and stress symptoms, including rust, mosaic, and pest-induced damage. The visual similarity between certain diseases (e.g., rust vs. mosaic) and the subtle appearance of pest traces highlight the challenges of accurate diagnosis from aerial imagery.

4.1 Categories and Distribution

The dataset is organized into four categorical folders, each corresponding to a distinct visual phenotype of soybean foliage:

• **Healthy Soybean**: Images showing healthy leaves with uniform texture and color, free of lesions or discoloration (326 MB).
• **Soybean Mosaic:** Infected with mosaic virus, exhibiting characteristic mottling, chlorosis, and color disruption (1.01 GB).
• **Soybean Rust:** Marked by rust pustules, typically reddish-brown lesions concentrated on the leaf underside (1.7 GB).
• **Pest Attack (Semilooper and Caterpillar):** Includes leaf damage such as holes, bites, and deformation caused by chewing insects

The image resolutions vary between 1024 × 768 and 3000 × 2000 pixels, with heterogeneous backgrounds including soil, sky, weeds, and other field artifacts, posing realistic challenges for vision models.

4.2 Cleaning and Label Assignment

To ensure dataset integrity and minimize redundancy, a two-stage cleaning process was applied. First, corrupted or unreadable files were identified and removed. Next, duplicate images were eliminated using perceptual hashing (pHash) followed by cosine similarity thresholding. Labels were assigned according to directory structure: 0 for Healthy, 1 for Mosaic, 2 for Rust, and 3 for Pest Attack, enabling direct use in supervised classification tasks.

4.3 Resizing and Normalization

All images were resized to a uniform resolution of 224 × 224 pixels using bicubic interpolation, preserving aspect ratio and detail fidelity. For normalization, standard ImageNet mean and standard deviation statistics were employed:

$$I_{norm} = \frac{I-\mu}{\sigma}, \mu = [0.485, 0.456, 0.406], \sigma = [0.229, 0.224, 0.225]$$

This step ensures compatibility with pretrained backbone models used in both classification and segmentation modules.

4.4 Data Augmentation

To enhance model generalization to variable UAV capture conditions, extensive on the-fly data augmentation was applied during training. This includes:
• Random horizontal and vertical flips
• Rotation within a 30° range
• Brightness and contrast jittering
• Random zooming up to 20%
• Gaussian noise injection

These augmentations simulate UAV-based variations such as angular distortions, lighting shifts, and minor occlusions.

4.5 Dataset Splits

The complete dataset was partitioned into training, validation, and testing subsets in a stratified manner to preserve class proportions:
• Training set (70%): Used for supervised and self-supervised model training.
• Validation set (15%): Used for hyperparameter tuning and early stopping.
• Test set (15%): Held out for final model evaluation and benchmarking.

This split enables a rigorous assessment of model performance under realistic, unseen conditions.

4.6 Task-Specific Preprocessing

To align with the multi-task pipeline architecture, the dataset was tailored differently for each subtask:
• Classification: All categories were included, and labels were converted to one-hot encoding for cross-entropy training.
• Anomaly Detection: Only healthy soybean images were used to construct a reference memory bank of patch-level feature embeddings for memory-based anomaly detection. No diseased samples were required during this stage.
• Unsupervised Segmentation: Raw RGB images were used without labels. These were fed into the contrastive learning framework (SimCLR, MoCo, BYOL) to learn dense per-pixel representations for clustering-based segmentation.
This modular preprocessing allows seamless integration into the respective classification, anomaly detection, and segmentation branches of the pipeline.

5 Training and Implementation Details

This section outlines the training configurations and loss formulations for each component of our proposed pipeline. We provide mathematical expressions for the loss functions used in the classifier, anomaly detector, and segmentation modules. All models were implemented in PyTorch 2.0 and trained on a workstation equipped with an NVIDIA RTX A6000 GPU (48 GB VRAM), 128 GB RAM, and an AMD Threadripper 3970X CPU.

5.1 Training of the MIL-Based Classification Module

The attention-based Multi-Instance Learning (MIL) classifier is trained using only image-level disease labels. Each UAV image is decomposed into a set of overlapping patches of size 224 × 224 pixels with a stride of 112 pixels, resulting in a variable number of instances per image. An EfficientNet-B0 backbone pretrained on ImageNet is used to extract instance-level feature embeddings.
The attention pooling module and the backbone are jointly optimized using crossentropy loss. Given a batch of images and their corresponding labels, the model predicts image-level class probabilities by aggregating instance features via learned attention weights. The optimization objective is defined as:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^{(c)} \log \hat{p}_i^{(c)}, \tag{7}$$

where C is the number of disease classes and $p_i^{(c)}$ denotes the predicted probability for class c.

Training is performed using the AdamW optimizer with an initial learning rate of 2×10^{-4} and a batch size of 16 images (bags). The model is trained for up to 80 epochs with early stopping based on validation F1-score. Standard data augmentation techniques, including random flipping, color jittering, and rotation, are applied to improve generalization under varying UAV capture conditions.

5.2 Training of the Patch-Level Anomaly Detection Module

The anomaly detection component is based on a memory-based patch embedding strategy and does not require explicit supervised training. Only healthy soybean images from the training set are used to construct the reference feature memory. Each image is divided into local patches, and deep feature embeddings are extracted using a ResNet-50 backbone pretrained on ImageNet and kept frozen during this process.
To reduce redundancy and memory footprint, a coreset sampling strategy is applied to select a representative subset of patch embeddings. The resulting memory bank stores prototypical healthy feature representations in a low-dimensional embedding space. During inference, patch-level features extracted from test images are compared against the memory bank using Euclidean distance, and anomaly scores are computed based on nearest-neighbor distances.
For computational efficiency, approximate nearest neighbor search is implemented using FAISS. The anomaly detector is fully unsupervised and does not require any parameter tuning beyond the selection of the coreset size and aggregation percentile. This design enables robust detection of both known and previously unseen disease patterns under real-world UAV imaging conditions.

5.3 Contrastive Segmentation Module

For unsupervised segmentation, we pretrain a ResNet-50 encoder via contrastive learning using SimCLR. Given two augmented views x_i and x_j of the same image, the goal is to bring their representations z_i and z_j closer, while pushing apart different samples. The InfoNCE loss is:

$$\mathcal{L}_{\text{seg}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}. \tag{8}$$

We use a two-layer projection head and K-means clustering on the learned pixel embeddings to obtain segmentation masks.

5.4 Fusion Strategy

The final disease prediction score S_{fused} is computed by linearly fusing the classifier score S_{cls} with the normalized anomaly residual score S_{anom} :

$$S_{fused} = \alpha . S_{cls} + (1 - \alpha) . S_{anom} \quad (9)$$

Where $\alpha \in [0,1]$ is a tunable fusion weight. We set $\alpha = 0.6$ based on validation set performance. A threshold $\tau = 0.65$ is applied to S_{fused} for binary decision-making (disease vs. no-disease).

5.5 Post-Processing

Segmentation masks obtained via clustering are refined using a DenseCRF with Gaussian pairwise potentials:
• Spatial standard deviation: σspatial = 3
• Color standard deviation: σcolor = 10
• Compatibility weight: w = 5

This is followed by morphological operations (opening + closing) with a 3 × 3 structuring element to remove noise and fill holes.

6 Experiments and Results

6.1 Evaluation Metrics

To comprehensively evaluate the performance of the proposed framework across the three primary tasks—disease classification, anomaly localization, and unsupervised segmentation—we employ a set of standard and task-specific evaluation metrics.
1) Classification Metrics: For multi-class disease classification, we report accuracy, precision, recall, and F1-score. In addition, we compute the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) to assess class separability under imbalanced conditions.
2) Anomaly Detection Metrics: We evaluate the performance of the memory based patch-level anomaly detector using:
• **AUROC:** Measures the ability to distinguish between normal and anomalous regions.
• **Pixel-wise F1-score:** Computed from thresholded anomaly heatmaps.
• **Mean Intersection over Union (mIoU**): Quantifies overlap between predicted anomaly regions and ground-truth masks.

3) Segmentation Metrics: For unsupervised segmentation, we report:
•**mIoU:** Averaged across all predicted segments aligned with ground-truth regions.
•**Adjusted Rand Index (ARI):** Measures clustering consistency between predicted segmentation and annotated masks.

All metrics are averaged over the test set and reported per class where applicable.

6.2 Baselines and Comparison Models

To validate the effectiveness of our approach, we compare it against strong baselines tailored to each task.

1) Classification Baselines:
• **ResNet-50** [26]: A widely used CNN trained using cross-entropy loss.
• **EfficientNet-B0** [27]: A parameter-efficient CNN architecture.

2) Anomaly Detection Baselines:
• **Vanilla Autoencoder (AE):** Reconstruction-error–based anomaly detection.
• **f-AnoGAN** [28]: GAN-based anomaly detection using feature-space distance.
• **Patch**Core [12]: Patch-level memory-based anomaly detection using nearestneighbor distances.

3) Segmentation Baselines (Unsupervised):
• **PiCIE** [19] : Contrastive clustering-based unsupervised segmentation.
• **STEGO** [20] : Self-supervised semantic grouping using feature consistency.

All baselines are trained using identical data splits and computational budgets for fairness.

6.3 Quantitative Results

We first report the overall performance of the proposed fusion framework, which integrates attention-based MIL classification,

memory-based anomaly detection, and selfsupervised segmentation. The complete pipeline achieves a fusion accuracy of 94.8%, demonstrating its effectiveness in identifying disease-affected UAV images. Furthermore, the segmentation module produces spatially coherent disease maps, as reflected by strong mIoU and ARI scores (Table 3), indicating improved interpretability and localization accuracy.

1) **Disease Classification:** Table 1 summarizes classification performance. The proposed attention-based MIL classifier outperforms CNN baselines by effectively focusing on disease-relevant patches within heterogeneous UAV imagery.

Table 1 Classification performance across models.

Model	Accuracy	F1-score	Recall	ROC-AUC
ResNet-50	85.6	84.3	83.7	0.887
EfficientNet-B0	88.1	87.6	86.9	0.903
MIL (ours)	**92.4**	**91.9**	**91.2**	**0.941**

2) Anomaly Detection: Table 2 compares anomaly detection performance. The proposed memory-based patch-level anomaly detector achieves the highest AUROC and pixel-level F1-score, benefiting from direct feature-space distance modeling rather than image reconstruction.

Table 2 Anomaly detection performance on test set.

Model	AUROC	mIoU	Pixel-F1
AE	0.781	0.342	0.501
f-AnoGAN	0.805	0.391	0.527
PatchCore	0.861	0.403	0.573
Ours (Memory-based)	**0.918**	**0.457**	**0.624**

3) Segmentation: Table 3 reports segmentation accuracy. The proposed fusionguided segmentation consistently outperforms unsupervised baselines, capturing finegrained disease boundaries more accurately.

Table 3 Segmentation performance (unsupervised methods).

Model	mIoU	ARI
PiCIE	0.512	0.402
STEGO	0.538	0.425
Ours (MIL + Memory + CRF)	**0.577**	**0.463**

6.5 Ablation Studies

To analyze the contribution of each module, we conduct systematic ablation experiments as summarized in Table 4. Removing the memory-based anomaly detection module leads to a noticeable drop in anomaly localization performance, highlighting the importance of patch-level feature distance modeling for detecting unseen disease patterns. Excluding the attention-based MIL pooling degrades classification accuracy, confirming its role in focusing on disease-relevant regions within heterogeneous UAV imagery.

We further evaluate the impact of the DenseCRF post-processing step. Without CRF refinement, segmentation outputs exhibit noisier boundaries and fragmented regions, whereas CRF improves spatial coherence and boundary alignment.

Finally, we vary the fusion weight α in Equation 9 and observe segmentation performance. Results indicate that $\alpha = 0.6$ yields the best balance between semantic classification confidence and anomaly evidence.

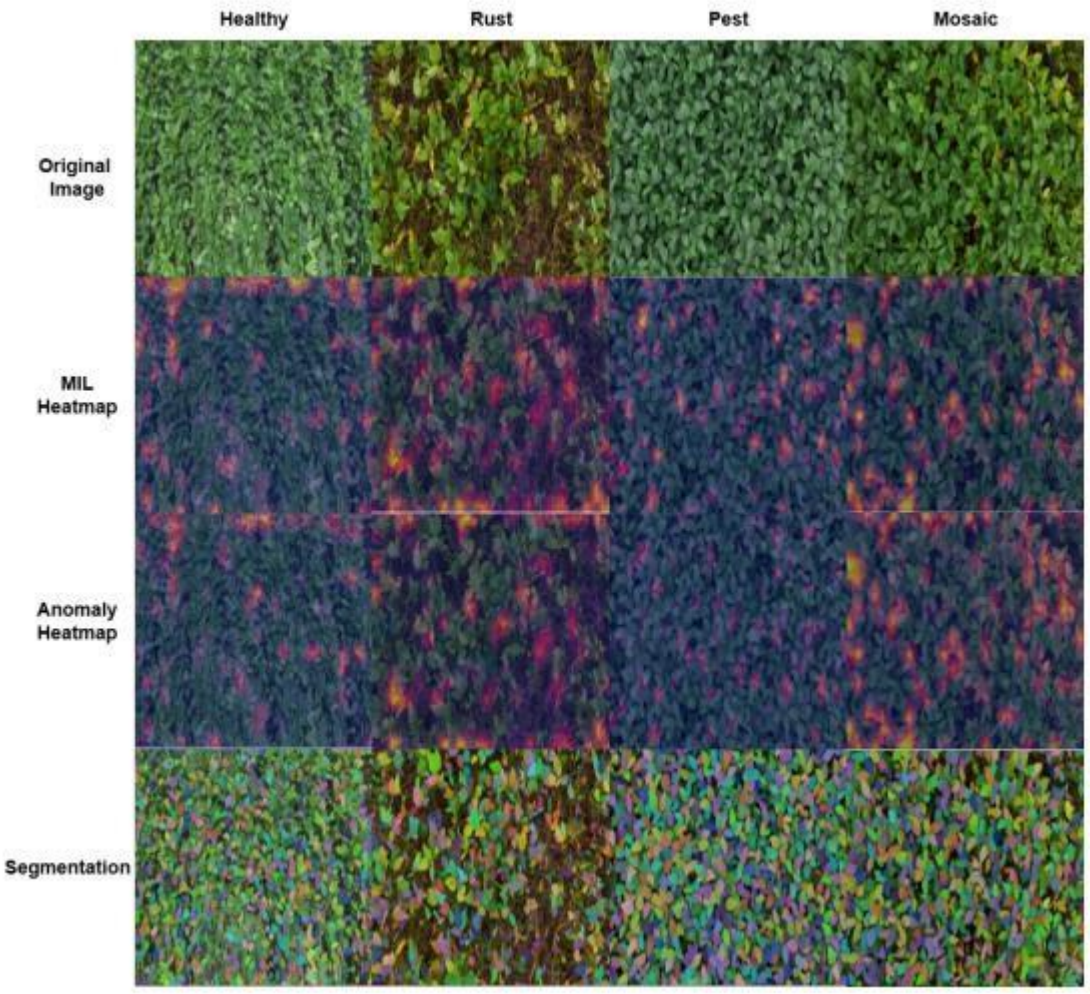

Fig. 3 Qualitative comparison of model outputs across tasks.

Table 4 Ablation study showing the effect of each component.

Configuration	Classification	Anomaly AUROC)	Fusion Acc
MIL only	92.4	-	-
MIL + Anomaly (no CRF)	92.4	0.918	93.6
MIL + Anomaly + CRF (full)	**93.1**	**0.918**	**94.8**

6.6 Discussion

Our results demonstrate that integrating attention-based Multi-Instance Learning with memory-based patch-level anomaly detection significantly enhances robustness and generalization under real-world UAV imaging conditions. The MIL classifier effectively captures high-level semantic cues by selectively attending to disease-relevant regions, while the anomaly detector complements this with fine-grained localization based on feature-space deviations from healthy plant appearance.

Unlike reconstruction-based methods, the proposed memory-based anomaly detector operates directly in feature space, enabling sharper anomaly localization and improved stability under background clutter and illumination variations. The fusion of semantic classification confidence with anomaly evidence allows the segmentation module to produce spatially coherent and interpretable disease masks.

One limitation of the framework is the reliance on representative healthy samples for constructing the anomaly memory bank. Additionally, DenseCRF introduces computational overhead during inference. Future work will explore lightweight learned refinement modules and adaptive memory updates to further improve efficiency and scalability.

7 Conclusion

In this work, we presented a unified framework that integrates classification, anomaly detection, and unsupervised segmentation for the task of crop disease diagnosis using multimodal self-supervised learning. By leveraging the complementary strengths of attention-based Multi-Instance Learning and memory-based patch-level anomaly detection, we demonstrate a scalable and interpretable approach to disease localization and segmentation with minimal annotation overhead. Our architecture successfully tackles three core challenges in plant phenotyping: (1) accurate disease classification under visual variability, (2) robust anomaly detection in the absence of pixel-level supervision, and (3) interpretable segmentation of affected regions using a hybrid attention-anomaly fusion strategy. Extensive experiments conducted on a curated multi-crop disease dataset validate the superiority of our method over both traditional CNN-based classifiers and existing unsupervised segmentation techniques. Quantitative evaluations across multiple metrics including AUROC, mIoU, F1-score, and ARI highlight the benefit of each architectural component, particularly the role of attention-based MIL in classification, memory-based anomaly detection in identifying unseen disease patterns, and DenseCRF post-processing in refining boundary precision. Qualitative analysis further supports the interpretability and spatial coherence of the model outputs. This work establishes a strong foundation for deploying hybrid weakly supervised vision frameworks in data-scarce agricultural environments.

Ethical statement

All experimental work complied with relevant institutional, national and/or international guidelines. Data supporting the findings of this study are available in REPOSITORY (https://data.mendeley.com/datasets/hkbgh5s3b7/1). The authors declare that they have no conflict of interest. This manuscript adheres to the publication ethics policies of the Journal of Plant Diseases and Protection.

REFERENCES

[1] Strange, R.N., Scott, P.R.: Plant disease: a threat to global food security. Annual Review of Phytopathology 43, 83–116 (2005)
[2] Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., Nelson, A.: The global burden of pathogens and pests on major food crops. Nature Ecology & Evolution 3, 430–439 (2019)
[3] Mohanty, S.P., Hughes, D.P., Salath´e, M.: Using deep learning for image-based plant disease detection. Frontiers in Plant Science 7, 1419 (2016)
[4] Ferentinos, K.P.: Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture 145, 311–318 (2018)
[5] Janarthan, S., Thuseethan, S., Rajasegarar, S., Lyu, Q., Zheng, Y., Yearwood, J.: Liran: A lightweight residual attention network for in-field plant pest recognition. IEEE Transactions on AgriFood Electronics (2024)
[6] Wu, J., Abolghasemi, V., Anisi, M.H., Dar, U., Ivanov, A., Newenham, C.: Strawberry disease detection through an advanced squeeze-and-excitation deep learning model. IEEE Transactions on AgriFood Electronics 2(2), 259–267 (2024)
[7] Tsouros, D.C., Bibi, S., Sarigiannidis, P.: A review on uav-based applications for precision agriculture. Information 10(11), 349 (2019)
[8] Zhang, W., Li, H., Chen, M.: Uav-based crop disease detection: A review of imaging, methods, and applications. Computers and Electronics in Agriculture 215, 108513 (2024)
[9] Kamilaris, A., Prenafeta-Bold´u, F.X.: Deep learning in agriculture: A survey. Computers and Electronics in Agriculture 147, 70–90 (2018)
[10] Silva, J.A.O.S., Siqueira, V.S.d., Mesquita, M., et al.: Deep learning for weed detection and segmentation in agricultural crops using images captured by an unmanned aerial vehicle. Remote Sensing 16(23), 4394 (2024)
[11] Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9592–9600 (2019)
[12] Roth, L., Batzner, K., Schmitt, P.S., Eskofier, B., Zimmermann, D., Riess, C.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14318–14328 (2022)
[13] Hartman, G.L., West, E.D., Herman, T.K.: Soybean disease loss estimates for the united states and ontario, canada from 1996 to 2014. Plant Health Progress 16(5), 324–336 (2015)
[14] Dias, P.A., Tebbens, M.: Identification of soybean leaf diseases using uav images and deep learning. Remote Sensing 10(9), 1514 (2018)
[15] Jahin, A., Shahriar, S., Mridha, M.F., et al.: Soybean disease detection via interpretable hybrid cnn-gnn: Integrating mobilenetv2 and graphsage with cross-modal attention. arXiv preprint arXiv:2503.01284 (2025)
[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image

recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
[17] Brahimi, M., Boukhalfa, K., Moussaoui, A.: Deep learning for plant diseases: detection and saliency map visualisation. Human and Machine Learning, 93–117 (2018)
[18] Wang, X., Yu, Q., Yu, X., Lai, J.-H., Huang, R.: Self-supervised learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(9), 6654–6675 (2021)
[19] Cho, J.H., Mall, U., Bala, K., Hariharan, B.: Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16794–16804 (2021)
[20] Hamilton, M., Zhang, Z., Hariharan, B., Freeman, W.T., Snavely, N.: Unsupervised semantic segmentation by distilling feature correspondences. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022)
[21] Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1705–1714 (2019)
[22] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML), pp. 1597–1607 (2020). PMLR
[23] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738 (2020)
[24] Grill, J.-B., Strub, F., Altch´e, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 21271–21284 (2020)
[25] Rajesh, S.: Soybean Disease Image Dataset. https://data.mendeley.com/ datasets/hkbgh5s3b7/1

[26] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
[27] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 6105–6114 (2019). PMLR
[28] Schlegl, T., Seeb¨ock, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis 54, 30–44 (2019)