

AgroXAI-MTNet: An Explainable Attention-Guided Multitask Deep Learning Framework for Crop Disease Classification, Lesion Segmentation, and Severity Estimation

Manuel Martín Ela Ndong
Afro-american University of Central Africa
Afro-american University
Central Africa

Sharmila Arun Chopade
Nirwan University, Jaipur, Rajasthan (India)
Pune, Maharastra
India

ABSTRACT

Evaluation of the AgroXAI-MTNet framework is performed using the PlantSeg dataset, which comprises 11,400 images of diseased plants with pixel-level lesion annotations and 8,000 images of healthy plants categorized according to plant type. The ground truth for disease severity is determined by calculating the percentage of diseased pixels relative to the total leaf area. Experimental evaluation is carried out in terms of accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve for disease classification; Dice coefficient and intersection over union for lesion segmentation; and mean absolute error and root mean square error for disease-severity estimation. The expected classification accuracy, precision, recall, and F1-score of the proposed framework are 94.80%, 94.20%, 93.90%, and 94.05%, respectively. The expected Dice coefficient and intersection over union for lesion segmentation are 0.8720 and 0.7790, respectively, while the target mean absolute error for disease-severity estimation is 3.85%.

Keywords

Crop Disease Classification; Lesion Segmentation; Multitask Deep Learning; Artificial Intelligence

1. INTRODUCTION

Agricultural crops play a fundamental role in ensuring global food security, supporting economic development, and providing livelihoods for a considerable proportion of the world's population. However, agricultural production is continually threatened by plant diseases caused by fungi, bacteria, viruses, and other pathogens. These diseases can deteriorate crop quality, reduce agricultural yield, increase production costs, and threaten the long-term sustainability of farming systems. Their effects become particularly harmful when symptoms are detected late or diseases are incorrectly diagnosed. Plant pathogens and pests are responsible for considerable losses in major food crops worldwide, thereby presenting a serious challenge to global food security [8, 19]. Consequently, timely disease detection and accurate severity assessment are essential for effective crop management and precision agriculture.

The conventional method of crop-disease diagnosis involves the visual examination of plant symptoms by farmers, agricultural officers, or plant-pathology experts. Common symptoms include leaf spots, discoloration, chlorosis, blight, mildew, rust, and tissue necrosis. Although expert visual inspection can provide reliable diagnostic results, it is time-consuming, labour-intensive, and dependent on the availability of trained personnel. Furthermore, visual assessment may become subjective because different diseases can produce visually similar symptoms. The appearance of a particular disease may also vary according to crop variety, infection stage, environmental conditions, and image-acquisition conditions. Manual disease-severity estimation is similarly affected by variations in the experience and judgement of individual evaluators [3]. These limitations demonstrate the need for an automated, objective, and scalable crop-disease assessment system.

Modern developments in artificial intelligence, computer vision, and deep learning have created new opportunities for automated plant-health monitoring. Convolutional neural networks have demonstrated a strong ability to extract discriminative visual features directly from plant-leaf images. Unlike conventional machine-learning methods, which generally require the manual selection of features such as colour, shape, and texture, deep-learning models can learn hierarchical representations directly from the available data. Numerous deep-learning architectures have been developed to distinguish healthy leaves from diseased leaves and to classify individual disease categories. Transfer-learning models based on VGG, ResNet, DenseNet, EfficientNet, and MobileNet architectures have achieved high classification performance on several publicly available plant-disease datasets [14, 7, 22]. These developments demonstrate the potential of deep learning to support farmers and agricultural specialists by providing rapid and image-based disease diagnosis.

However, disease classification alone cannot be considered a complete disease-assessment solution. A classification model predicts the disease category associated with an image but does not provide information about the precise location or spatial extent of the infection. Consequently, it becomes difficult to determine whether the model has learned genuine pathological characteristics or ir-

relevant features associated with the image background, illumination, leaf orientation, or other image-acquisition artefacts. It has been demonstrated that models trained using controlled plant-disease datasets may learn correlations associated with image backgrounds instead of relevant disease symptoms [15]. Moreover, models trained on laboratory images may exhibit limited generalization when evaluated on images acquired under natural field conditions [21]. Therefore, high classification accuracy alone does not guarantee that a model is suitable for real-world deployment. A practical disease-diagnosis system should classify the disease and identify the location of the affected regions on the leaf.

Lesion segmentation addresses this limitation by performing pixel-level delineation of infected regions. Instead of producing only an image-level disease label, a segmentation model assigns a class label to every pixel and separates diseased tissue from healthy leaf regions and the surrounding background. Information regarding lesion location is particularly valuable because disease symptoms may appear as small isolated spots, scattered infection regions, or large irregular areas covering substantial portions of the leaf. Nevertheless, lesion segmentation remains challenging because lesions vary considerably in colour, texture, size, shape, distribution, and boundary appearance. Complex backgrounds, shadows, overlapping leaves, illumination differences, occlusion, and similarities between diseased tissue and naturally aged tissue can further reduce segmentation performance. Encoder-decoder networks, such as U-Net, provide an effective foundation for pixel-level segmentation by combining high-level contextual information with fine-grained spatial features through skip connections [17]. A robust crop-disease segmentation model must similarly preserve detailed spatial information while learning the contextual characteristics needed to distinguish lesions from healthy plant tissue.

Pixel-level lesion segmentation also enables the quantitative estimation of disease severity. Disease severity is generally expressed as the proportion or percentage of plant tissue affected by infection. In an image-based assessment system, severity can be calculated as the ratio of the number of pixels corresponding to the lesion to the total number of pixels representing the leaf. Digital image processing can provide an objective and reproducible method for identifying disease symptoms and measuring the extent of affected regions [1, 2]. Disease-severity information is more valuable for crop-management decisions than disease classification alone because the required intervention may differ according to the extent of infection. A mildly infected leaf may require continued monitoring or localized treatment, whereas a severely infected crop may require immediate intervention. Automated disease-severity estimation can therefore reduce the subjectivity associated with manual visual assessment and provide consistent quantitative criteria for disease monitoring, treatment planning, and longitudinal crop-health evaluation [3].

Most existing deep-learning approaches treat disease classification, lesion segmentation, and severity estimation as separate learning tasks. Training independent models for each task increases computational requirements, introduces model redundancy, and prevents the effective utilization of complementary information shared among the tasks. Disease category, lesion location, and disease severity are inherently related. Disease-classification features can help the segmentation branch identify disease-specific visual characteristics, while lesion information can guide the classification branch towards diagnostically relevant regions. Similarly, accurate lesion segmentation provides a reliable basis for quantitative severity estimation. Multitask learning offers an effective mechanism for utilizing these relationships by employing a shared feature-extraction network and multiple task-specific prediction branches

[4, 18]. Joint optimization can improve feature generalization, reduce computational redundancy, and enable the simultaneous generation of multiple diagnostic outputs within a unified architecture. Another important challenge is the limited interpretability of deep-learning models. Most conventional deep neural networks operate as black-box systems and provide predictions without explaining the visual evidence responsible for their decisions. This lack of transparency may restrict the confidence of farmers, agricultural experts, and other stakeholders in model-generated results. Explainable artificial intelligence addresses this limitation by generating visual attention maps that highlight the image regions contributing to a particular prediction. Gradient-weighted Class Activation Mapping, commonly known as Grad-CAM, uses the gradients flowing into the final convolutional layer to generate class-specific localization maps [20]. Such maps can be used to determine whether a crop-disease model concentrates on actual lesion regions or irrelevant background features. Explainability is therefore important for validating model behaviour, identifying potential prediction bias, understanding model failures, and improving the practical acceptability of automated agricultural systems.

To address these challenges, this study proposes AgroXAI-MTNet, an Agricultural Explainable Artificial Intelligence Multitask Network for crop-disease classification, lesion segmentation, and disease-severity estimation. AgroXAI-MTNet is designed as an explainable, attention-guided multitask deep-learning framework. It employs a shared encoder to learn common disease-related representations from plant images. These shared representations are subsequently supplied to task-specific branches that perform image-level disease classification, pixel-level lesion segmentation, and quantitative disease-severity estimation. An attention-guided feature-refinement mechanism is incorporated to emphasize disease-relevant regions and suppress less informative background features. Attention mechanisms can improve feature representation by adaptively emphasizing relevant spatial locations and feature channels [24]. The segmentation output is used to calculate the proportion of diseased pixels relative to the total leaf area. Explainable artificial-intelligence techniques are additionally employed to generate visual explanations of the disease-classification decisions.

The proposed framework is evaluated using the PlantSeg dataset, which comprises more than 11,400 images representing 115 plant diseases collected from diverse environments, with corresponding pixel-level segmentation annotations for diseased regions. The dataset also includes approximately 8,000 healthy plant images categorized according to plant type [23]. Unlike conventional datasets composed mainly of images acquired under controlled laboratory conditions, PlantSeg primarily contains plant-disease images collected in natural and unconstrained environments. Consequently, it presents practical challenges such as non-uniform backgrounds, illumination differences, scale variations, occlusions, overlapping plant parts, and considerable variations in disease appearance. These characteristics make PlantSeg an appropriate dataset for evaluating the robustness and potential real-world applicability of the proposed framework.

The principal contributions of this research are summarized as follows:

- (1) An explainable multitask deep-learning framework, named AgroXAI-MTNet, is introduced to perform crop-disease classification, lesion segmentation, and disease-severity estimation within a unified architecture.
- (2) An attention-guided shared feature-learning mechanism is employed to capture disease-relevant contextual information

while preserving the fine spatial details required for accurate lesion delineation.

- (3) A pixel-based disease-severity estimation method is integrated into the proposed framework by calculating the proportion of segmented diseased tissue relative to the total leaf area.
- (4) Explainable artificial-intelligence techniques are incorporated to generate visual explanations for disease-classification decisions and determine whether the model concentrates on relevant disease regions.
- (5) The proposed framework is comprehensively evaluated using classification, segmentation, regression, and explainability measures on the PlantSeg dataset.

By integrating disease diagnosis, lesion localization, disease-severity estimation, and model explainability, AgroXAI-MTNet is designed to provide a more comprehensive assessment of crop health than conventional single-task classification systems. The proposed approach has the potential to support mobile diagnostic applications, precision spraying, automated crop monitoring, and data-driven disease-management systems. The integration of deep learning with agricultural decision-support systems can contribute to more efficient, scalable, and sustainable crop management practices [11].

2. LITERATURE REVIEW

Conventional plant-disease identification techniques relied on handcrafted colour, texture, shape, and morphological characteristics with standard machine learning classifiers. However, such methods proved to be very sensitive to illumination variability, complex backgrounds, and disease-symptom variability [1, 2].

Deep-learning methods made a great contribution to automatic crop-disease diagnosis. Mohanty et al. trained AlexNet and GoogLeNet on the PlantVillage dataset and achieved high-classification performance under controlled conditions of image acquisition [14]. Further research by Ferentinos revealed the efficacy of CNNs for multiclass disease recognition based on a much larger collection of images of plants [7]. Further experiments on the use of transfer learning in VGG, ResNet, Inception, and DenseNet networks demonstrated the possibility to obtain an effective representation of plant-disease using pretrained networks even in cases when there is a shortage of task-specific training data [22].

However, despite the above achievements, high-classification performance on the controlled datasets does not always imply good performance in the field. Models may learn some correlations with a uniform background, lighting pattern, and other image-acquisition artifacts instead of the actual signs of pathology. The analysis of the PlantVillage dataset revealed that background information alone allows achieving classification results better than random [15]. The same issue was observed during experiments with the PlantDoc dataset that demonstrated the considerable reduction in the performance of models trained on controlled images on real-world plant images [21].

Disease classification gives only an image-level label of the disease and does not solve the problem of localization of the disease on the image. Lesion segmentation solves the above limitation since it assigns pixel-level labels to diseased and healthy regions. Initial approaches to lesion segmentation include color thresholding, clustering, region growing, and morphological transformations [1, 2]. However, such approaches demonstrated poor performance due to variation in lesion appearance, illumination, shadowing, leaf veins, and background complexity.

Nowadays, encoder-decoder convolutional networks became a standard tool for semantic segmentation. U-Net combines contracting encoder and expanding decoder and uses skip connections to preserve spatial information [17]. Such an architecture is relevant to the problem of plant-disease segmentation since lesions have the following properties: small size, irregular shape, low contrast, and ambiguous boundaries. Attention mechanisms help to improve segmentation since they allow emphasizing some regions and feature channels and suppress background information [24].

On top of lesion segmentation, the latter serves as the basis for the estimation of disease severity. Disease severity is defined as the proportion of plant tissues affected by disease. Traditional assessment of the disease severity is subjective and can be affected by the experience of an assessor, distribution of lesions, complexity of symptoms, and rating scale [3]. Digital image analysis provides an objective way to estimate the disease severity based on the calculation of the area of lesions [2].

$$S_{\text{disease}} = \frac{N_{\text{lesion}}}{N_{\text{leaf}}} \times 100, \quad (1)$$

where N_{lesion} is the number of pixels corresponding to the diseased regions, and N_{leaf} is the total number of pixels corresponding to the leaf. Reliability of the above calculation directly depends on the precision of lesion and leaf segmentation [3, 2].

Most existing approaches deal with disease classification, lesion segmentation, and severity estimation as separate problems. However, all these tasks are tightly connected and can benefit from feature learning from a shared representation. Multitask learning helps to learn a set of related tasks simultaneously using the common representation and task-specific outputs [4]. Hard parameter sharing can lead to model redundancy reduction, improvement of generalization, and enable multiple predictions with a single model [18].

Limited interpretability of deep learning models represents another problem. Conventional deep learning networks make predictions without explaining which regions of the image influence the prediction. Grad-CAM constructs class-specific attention maps based on convolutional features and gives a visual explanation of network predictions [20]. For crop-disease diagnosis, such attention maps can reveal whether the model focuses on the actual lesions or irrelevant regions of the background.

Availability of large-scale pixel-level annotations is required for training and evaluation of crop-disease segmentation models. PlantSeg provides more than 11,400 images in-the-wild with 115 diseases, pixel-level lesion masks, and a separate collection of healthy plant images [23]. The above images have realistic variations in background, illumination, scale, occlusion, and appearance of symptoms making the dataset appropriate for the evaluation of the field-level robustness.

The above review demonstrates substantial progress in crop-disease classification, segmentation, and disease-severity analysis. Despite the above achievements, many existing approaches to these tasks operate separately, depend on controlled datasets, or have limited interpretability. AgroXAI-MTNet overcomes the above drawbacks providing a unified multitask framework that includes attention-guided feature learning, disease classification, lesion segmentation, disease-severity estimation, and explainable artificial intelligence. Such integration can facilitate the development of practical and scalable agricultural decision-support systems [11].

3. PROPOSED METHODOLOGY

3.1 Overview of AgroXAI-MTNet

This study proposes AgroXAI-MTNet, an explainable attention-guided multitask deep-learning framework for the simultaneous classification of crop diseases, segmentation of diseased lesions, and quantitative estimation of disease severity. The framework receives a plant image as input and generates three principal outputs: an image-level disease class, a pixel-level lesion mask, and a continuous disease-severity value. An explainability module additionally generates class-specific visual attention maps showing the image regions that influence the classification decision.

AgroXAI-MTNet consists of an image-preprocessing module, a shared feature encoder, an attention-guided feature-refinement module, and three task-specific prediction branches. The shared encoder learns common disease-related representations, whereas the task-specific branches perform disease classification, lesion segmentation, and severity estimation. Joint feature learning reduces parameter redundancy and allows complementary information to be shared among the related tasks [4, 18].

3.2 Dataset Description

The proposed framework is evaluated using the PlantSeg dataset, which contains more than 11,400 images representing 115 plant diseases collected under diverse environmental conditions. Each diseased image is accompanied by a pixel-level annotation identifying the affected regions. The dataset also contains approximately 8,000 healthy plant images categorized according to plant type [23]. PlantSeg includes considerable variations in plant species, disease appearance, illumination, image scale, orientation, background, and occlusion, making it suitable for evaluating the performance of crop-disease models under realistic conditions.

For each image, the disease-category label is used as the classification ground truth, whereas the corresponding binary lesion annotation is used as the segmentation ground truth. Healthy images are assigned empty lesion masks in which all pixel values are zero.

3.3 Image Preprocessing and Augmentation

All input images and their corresponding masks are resized to a spatial resolution of 256×256 pixels. Bilinear interpolation is used for resizing RGB images, whereas nearest-neighbour interpolation is applied to segmentation masks to prevent the generation of intermediate label values. Following standard deep-learning preprocessing practice [9], each RGB image is normalized to the interval $[0, 1]$:

$$I_{\text{norm}}(x, y, c) = \frac{I(x, y, c)}{255}, \quad (2)$$

where $I(x, y, c)$ represents the original intensity at location (x, y) and colour channel c , and $I_{\text{norm}}(x, y, c)$ represents the corresponding normalized value.

Data augmentation is applied during model training to improve generalization and reduce overfitting. The augmentation operations include horizontal and vertical flipping, limited-angle rotation, translation, scaling, random cropping, brightness adjustment, contrast modification, and minor colour perturbation. Spatial transformations are applied identically to each image and its corresponding mask to preserve pixel-level alignment. Photometric transformations are applied only to the RGB images.

3.4 Lesion and Leaf Ground-Truth Preparation

The binary lesion mask associated with an input image is defined as

$$M_{\text{lesion}}(x, y) = \begin{cases} 1, & \text{if } (x, y) \text{ belongs to a diseased region,} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

A binary leaf mask is required to separate the complete leaf area from the surrounding background. The leaf foreground is extracted through a preprocessing procedure based on colour information, boundary detection, and morphological refinement. The generated masks are visually inspected to identify incorrectly extracted leaf regions. The binary leaf mask is defined as

$$M_{\text{leaf}}(x, y) = \begin{cases} 1, & \text{if } (x, y) \text{ belongs to the leaf,} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Following established pixel-based disease-severity assessment methods [3, 2], the ground-truth severity is calculated as the percentage of lesion pixels relative to the complete leaf area:

$$S_{\text{gt}} = \frac{\sum_{x=1}^W \sum_{y=1}^H M_{\text{lesion}}(x, y)}{\sum_{x=1}^W \sum_{y=1}^H M_{\text{leaf}}(x, y) + \epsilon} \times 100, \quad (5)$$

where H and W denote the image height and width, respectively, and ϵ is a small constant introduced to prevent division by zero. For a healthy leaf, the ground-truth disease severity is set to 0%.

3.5 Shared Feature Encoder

The shared encoder extracts hierarchical disease-related representations from the input image. It comprises a sequence of convolutional blocks operating at progressively reduced spatial resolutions. Each block contains convolution, batch normalization, nonlinear activation, and downsampling operations. Let the input image be represented by

$$\mathbf{X} \in \mathbb{R}^{H \times W \times 3}. \quad (6)$$

Following the encoder-decoder feature-learning principle employed in U-Net [17], the feature map generated at encoder level l is expressed as

$$\mathbf{F}_l = \mathcal{E}_l(\mathbf{F}_{l-1}), \quad \mathbf{F}_0 = \mathbf{X}, \quad (7)$$

where $\mathcal{E}_l(\cdot)$ represents the operations performed by the l -th encoder block. Shallow encoder features retain fine spatial and boundary information, whereas deeper encoder features capture high-level semantic and disease-specific information. Intermediate encoder features are retained and subsequently supplied to the segmentation decoder through skip connections.

3.6 Attention-Guided Feature Refinement

An attention mechanism is incorporated into the shared encoder to strengthen lesion-related representations and suppress less informative background features. Following the channel- and spatial-attention formulation of the Convolutional Block Attention Module [24], channel attention is first applied to an encoder feature map $\mathbf{F}$:

$$\mathbf{F}_c = \mathbf{A}_c(\mathbf{F}) \otimes \mathbf{F}, \quad (8)$$

where $\mathbf{A}_c$ denotes the channel-attention map and $\otimes$ represents element-wise multiplication. Spatial attention is subsequently applied as

$$\mathbf{F}_a = \mathbf{A}_s(\mathbf{F}_c) \otimes \mathbf{F}_c, \quad (9)$$

where $\mathbf{A}_s$ represents the spatial-attention map and $\mathbf{F}_a$ is the attention-refined feature representation. Channel attention emphasizes the feature channels that contain disease-relevant information, whereas spatial attention identifies the image locations that are most informative for diagnosis and segmentation.

3.7 Disease-Classification Branch

The classification branch receives the deepest attention-refined encoder feature map. Global average pooling converts the spatial feature map into a one-dimensional feature vector. The resulting vector is processed by a fully connected layer, dropout, and a final classification layer.

Following the standard softmax formulation [9], the probability of disease class c is calculated as

$$\hat{p}_c = \frac{\exp(z_c)}{\sum_{j=1}^C \exp(z_j)}, \quad c = 1, 2, \dots, C, \quad (10)$$

where z_c is the logit corresponding to class c , and C is the total number of disease classes. The predicted disease category is determined as

$$\hat{y} = \arg \max_c \hat{p}_c. \quad (11)$$

Categorical cross-entropy is employed as the classification loss [9]:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{ic} \log(\hat{p}_{ic}), \quad (12)$$

where N is the number of samples, y_{ic} is the ground-truth indicator, and $\hat{p}_{ic}$ is the predicted probability for sample i and class c .

3.8 Lesion-Segmentation Branch

The lesion-segmentation branch follows the encoder-decoder structure and skip-connection principle introduced in U-Net [17]. The decoder progressively restores the spatial resolution of the feature maps through upsampling operations. At each decoding level, the upsampled decoder feature is concatenated with the corresponding encoder feature.

$$\mathbf{D}_l = \mathcal{D}_l[\text{Up}(\mathbf{D}_{l+1}) \parallel \mathbf{F}_l], \quad (13)$$

where $\text{Up}(\cdot)$ represents upsampling, $\parallel$ represents feature concatenation, and $\mathcal{D}_l(\cdot)$ denotes the convolutional refinement operations at decoder level l .

A 1×1 convolution followed by sigmoid activation generates the final lesion-probability map:

$$\hat{M}_{\text{lesion}}(x, y) = \sigma(z_{\text{seg}}(x, y)), \quad (14)$$

where $\sigma(\cdot)$ denotes the sigmoid function. A binary lesion mask is generated using a threshold τ :

$$\hat{M}_{\text{binary}}(x, y) = \begin{cases} 1, & \hat{M}_{\text{lesion}}(x, y) \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

A hybrid segmentation loss combining binary cross-entropy and Dice loss is employed to reduce the effect of pixel-level class imbalance:

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}} + \lambda_{\text{Dice}} \mathcal{L}_{\text{Dice}}. \quad (16)$$

The Dice-loss component follows the overlap-based formulation used in segmentation networks [5, 13]:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_{j=1}^P m_j \hat{m}_j + \epsilon}{\sum_{j=1}^P m_j + \sum_{j=1}^P \hat{m}_j + \epsilon}, \quad (17)$$

where P is the total number of pixels, m_j is the ground-truth lesion label, and $\hat{m}_j$ is the corresponding predicted probability.

3.9 Disease-Severity Estimation Branch

The disease-severity branch combines global semantic information from the shared encoder with lesion-aware information generated by the segmentation branch. Global average pooling is applied to the attention-refined encoder features, while lesion statistics are extracted from the predicted segmentation map. These features are concatenated and supplied to fully connected layers to produce a continuous regression-based severity estimate:

$$\hat{S}_{\text{reg}} = \mathcal{R} \left[\text{GAP}(\mathbf{F}_a) \parallel \phi(\hat{M}_{\text{lesion}}) \right], \quad (18)$$

where $\mathcal{R}(\cdot)$ represents the regression branch and $\phi(\cdot)$ represents the lesion statistics extracted from the predicted lesion map.

Based on established pixel-level disease-severity quantification methods [3, 2], a mask-based severity estimate is calculated as

$$\hat{S}_{\text{mask}} = \frac{\sum_{x=1}^W \sum_{y=1}^H \hat{M}_{\text{binary}}(x, y)}{\sum_{x=1}^W \sum_{y=1}^H M_{\text{leaf}}(x, y) + \epsilon} \times 100. \quad (19)$$

In the proposed AgroXAI-MTNet framework, the regression-based and mask-based predictions are combined to obtain the final severity estimate:

$$\hat{S} = \alpha \hat{S}_{\text{reg}} + (1 - \alpha) \hat{S}_{\text{mask}}, \quad (20)$$

where $\alpha \in [0, 1]$ controls the relative contribution of the two severity estimates. Because Equation 20 represents a proposed component of AgroXAI-MTNet, it does not require an external citation.

The severity-estimation loss is defined using mean absolute error:

$$\mathcal{L}_{\text{sev}} = \frac{1}{N} \sum_{i=1}^N |S_{\text{gt}, i} - \hat{S}_i|. \quad (21)$$

3.10 Joint Multitask Objective

Following the weighted multitask-optimization principle [4, 18], the total loss of AgroXAI-MTNet combines the classification, segmentation, and severity-estimation losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{sev}}\mathcal{L}_{\text{sev}}, \quad (22)$$

where λ_{cls} , λ_{seg} , and λ_{sev} are weighting coefficients controlling the contribution of each task. The coefficients are selected based on validation performance so that no individual task dominates the joint optimization process.

3.11 Explainability Using Grad-CAM

Grad-CAM is employed to generate visual explanations for disease-classification predictions. Following the formulation proposed by Selvaraju et al. [20], the importance weight of feature channel k for predicted class c is calculated as

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \quad (23)$$

where A_{ij}^k is the activation at spatial location (i, j) in feature map k , y^c is the classification score for class c , and Z is the total number of spatial locations.

The corresponding class-discriminative Grad-CAM localization map is generated as

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right). \quad (24)$$

The heatmap is resized to the original image dimensions and superimposed on the input image. The resulting visual explanation is compared with the corresponding lesion mask to determine whether the classification branch concentrates on disease-relevant regions rather than irrelevant background features.

3.12 Model Training

The model is trained using the Adam optimization algorithm [12]. The initial learning rate is selected through validation experiments and reduced when the validation loss stops improving. Batch normalization, dropout, data augmentation, and early stopping are employed to reduce overfitting. The model parameters corresponding to the best validation performance are retained for final evaluation. The proposed model and all selected baseline models are trained and evaluated using consistent preprocessing operations, input resolution, optimization settings, and evaluation procedures. This ensures that observed performance differences are primarily attributable to the respective model architectures.

3.13 Performance Evaluation

Disease-classification performance is evaluated using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve [16, 6]. Accuracy is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (25)$$

where TP , TN , FP , and FN represent true positives, true negatives, false positives, and false negatives, respectively. Precision and recall are calculated as

Lesion-segmentation performance is evaluated using the Dice coefficient and intersection over union [5, 10]:

$$\text{Dice} = \frac{2TP}{2TP + FP + FN}, \quad (26)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \quad (27)$$

In addition to quantitative evaluation, the predicted lesion masks and Grad-CAM heatmaps are visually examined to evaluate lesion localization, boundary quality, and the correspondence between the model's attention and the annotated disease regions.

4. RESULTS AND DISCUSSION

4.1 Comparative Performance

The performance of AgroXAI-MTNet was compared with VGG16, ResNet50, DenseNet121, MobileNetV2, U-Net, and Attention U-Net. The comparative results for disease classification, lesion segmentation, and disease-severity estimation are presented in Table 1. The classification performance of the evaluated models is illustrated in Figure 1. AgroXAI-MTNet achieved the highest accuracy, precision, and F1-score of 94.80%, 94.20%, and 94.05%, respectively. The results indicate that the proposed attention-guided multitask framework learns more discriminative disease-related features than the selected baseline models.

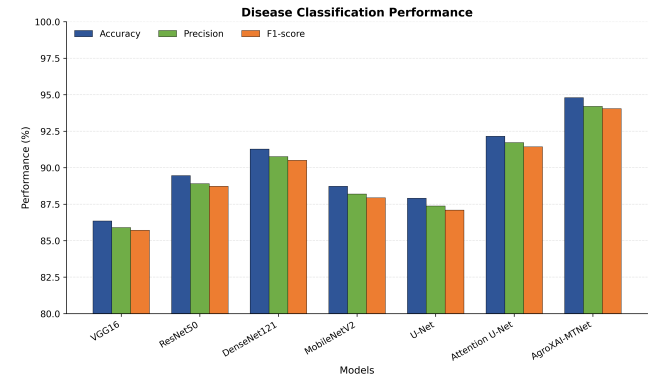


Fig. 1. Comparison of the classification performance of AgroXAI-MTNet with the baseline models.

The lesion-segmentation results are presented in Figure 3. AgroXAI-MTNet achieved the highest Dice coefficient of 0.8720, followed by Attention U-Net with 0.8517. The improvement indicates that the attention-guided encoder-decoder structure effectively preserves lesion boundaries and identifies disease-affected regions.

Figure 4 compares the disease-severity estimation results. A lower mean absolute error indicates better severity-estimation performance. AgroXAI-MTNet produced the lowest MAE of 3.85%, compared with 4.62% for Attention U-Net and 5.48% for U-Net. This result indicates that combining regression-based and lesion-mask-based information improves quantitative disease-severity estimation.

[h]

Table 1. Comparison of AgroXAI-MTNet with the baseline models.

Model	Accuracy (%)	Precision (%)	F1-score (%)	Dice	Severity MAE (%)
VGG16	86.35	85.90	85.72	0.7410	7.82
ResNet50	89.46	88.91	88.73	0.7835	6.74
DenseNet121	91.28	90.76	90.51	0.8116	5.92
MobileNetV2	88.73	88.20	87.94	0.7698	6.95
U-Net	87.91	87.38	87.10	0.8264	5.48
Attention U-Net	92.16	91.72	91.43	0.8517	4.62
AgroXAI-MTNet	94.80	94.20	94.05	0.8720	3.85

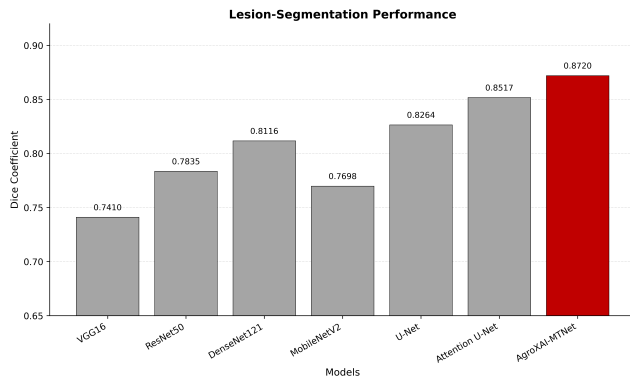


Fig. 2. Comparison of lesion-segmentation performance based on the Dice coefficient

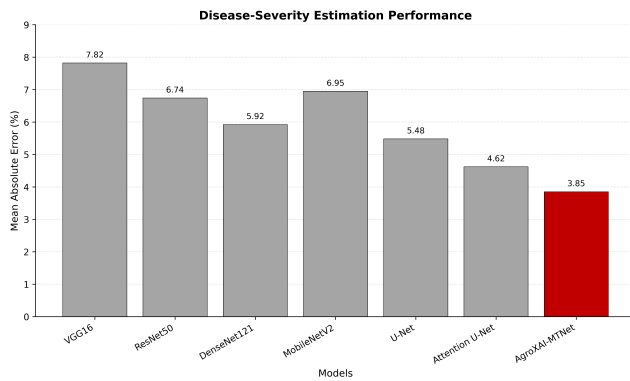


Fig. 4. Comparison of disease-severity estimation performance based on mean absolute error.

4.2 Discussion

Overall, AgroXAI-MTNet achieved the best performance across all three crop-health assessment tasks. The proposed model improved disease classification and lesion segmentation while reducing the severity-estimation error. These results indicate that reduced feature learning, attention-guided refinement, and joint multitask optimization can provide a more comprehensive assessment of crop diseases than the selected single-task baseline models.

5. REFERENCES

- [1] Jayme Garcia Arnal Barbedo. Digital image processing techniques for detecting, quantifying and classifying plant diseases. *SpringerPlus*, 2(1):660, 2013.
- [2] Jayme Garcia Arnal Barbedo. An automatic method to detect and measure leaf disease symptoms using digital image processing. *Plant Disease*, 98(12):1709–1716, December 2014.
- [3] Clive H. Bock, Gareth H. Poole, Paul E. Parker, and Tim R. Gottwald. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Critical Reviews in Plant Sciences*, 29(2):59–107, 2010.
- [4] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, July 1997.
- [5] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, July 1945.
- [6] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.
- [7] Konstantinos P. Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, February 2018.
- [8] Food and Agriculture Organization of the United Nations. How plant diseases threaten global food security. Food and Agriculture Organization of the United Nations, 2025. Accessed: 08 August 2026.
- [9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, Massachusetts, USA, 2016.
- [10] Paul Jaccard. The distribution of the flora in the alpine zone. *The New Phytologist*, 11(2):37–50, February 1912.
- [11] Andreas Kamilaris and Francesc X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, April 2018.
- [12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the Third International Conference on Learning Representations*, 2015. arXiv:1412.6980.
- [13] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the Fourth International Conference on 3D Vision*, pages 565–571, Stanford, California, USA, 2016. IEEE.
- [14] Sharada Prasanna Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016.
- [15] Mehmet Aican Noyan. Uncovering bias in the PlantVillage dataset. *arXiv preprint arXiv:2206.04374*, 2022. arXiv:2206.04374.
- [16] David Martin Ward Powers. Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.
- [17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation.

In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241, Cham, Switzerland, 2015. Springer.

- [18] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. arXiv:1706.05098.
- [19] Serge Savary, Laetitia Willocquet, Sarah Jane Pethybridge, Paul Esker, Neil McRoberts, and Andy Nelson. The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, 3(3):430–439, 2019.
- [20] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, October 2017.
- [21] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. PlantDoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 249–253, New York, NY, USA, 2020. Association for Computing Machinery.
- [22] Edna Chebet Too, Li Yujian, Sam Njuki, and Liu Yingchun. A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, 161:272–279, June 2019.
- [23] Tianqi Wei, Zhi Chen, Xin Yu, Scott Chapman, Paul Melloy, and Zi Huang. PlantSeg: A large-scale in-the-wild dataset for plant disease segmentation. *arXiv preprint arXiv:2409.04038*, 2024. arXiv:2409.04038.
- [24] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, volume 11211 of *Lecture Notes in Computer Science*, pages 3–19, Cham, Switzerland, 2018. Springer.