agriculture

MDPI

*Article*

# Fungal Disease Detection on Vegetable Crops Using RGB Imaging

**Charalampos Rafail Medentzidis, Dimitrios Kapetas and Eleftheria Maria Pechlivani ***

Centre for Research and Technology Hellas, Information Technologies Institute, 57001 Thessaloniki, Greece; cmedentzidis@iti.gr (C.R.M.); dimikape@iti.gr (D.K.)
* Correspondence: riapechl@iti.gr; Tel.: +30-231-125-7751

**Abstract**

The fungal pathogen *Botrytis cinerea* (*B. cinerea*) attacks over 1400 plant species and results in estimated annual losses of $10–100 billion worldwide. In precision agriculture, deep learning (DL) provides reliable tools for rapid and objective plant disease detection. This study presents a unified two-stage DL solution for the automated detection of visible *B. cinerea* across three major vegetable crops—tomato, pepper, and cucumber—using standard RGB imagery. In the first stage, a YOLOv11-based instance segmentation model accurately localizes leaf regions, achieving a localization accuracy of 87.3% as measured by $mAP_{50}$. In the second stage, an ensemble of 13 MobileViT variant models analyzes the segmented leaf regions and performs per-crop classification into healthy and infected leaves. The proposed system achieves an overall detection accuracy of 84.05%, with per-class detection of infected leaves at 88.61% for pepper, 82.68% for tomato, and 70.55% for cucumber, measured using the F1-score. These results demonstrate that the proposed approach can reliably detect *B. cinerea* symptoms across different crops using only RGB data, offering a practical path toward smartphone-based field deployment and integration into decision support systems for timely, symptom-based disease management.

**Keywords:** *Botrytis cinerea*; deep learning; YOLO; MobileViT; instance segmentation; ensemble learning; precision agriculture; computer vision; plant disease detection; RGB imaging

## 1. Introduction

Tomatoes, peppers, and cucumbers are among the most widely cultivated vegetables worldwide, prized for their nutritional value, culinary versatility, and economic importance. According to FAO data, tomato production reached approximately 186 million tonnes in 2022, while cucumber production was about 95 million tonnes [1]. Peppers also reached a record production of around 37 million tonnes globally in the same year [2]. Despite their extensive cultivation, these crops remain highly susceptible to biotic and abiotic stresses that can drastically reduce both yield and quality. Among the biotic stresses affecting these crops, the fungal pathogen *B. cinerea* poses severe economic and management challenges [3].

*B. cinerea* is one of the most devastating fungal pathogens affecting high-value crops, infecting over 1400 plant species and causing estimated annual economic losses up to $100 billion [4]. The disease spreads rapidly under Mediterranean climate conditions, particularly elevated humidity and moderate temperatures [5], and can affect the entire agricultural supply chain, from field to post-harvest storage and transport [6]. Moreover, the increasing resistance of B. cinerea to common fungicides [7] makes chemical treatments less effective and highlights the need for improved detection and management strategies.

Check for updates

Academic Editor: Jiehao Li

Received: 22 January 2026
Revised: 20 February 2026
Accepted: 25 February 2026
Published: 27 February 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Traditionally, diagnosis relies on visual inspection and laboratory methods [8], which are labor-intensive, time-consuming, and prone to human error. Therefore, accurate identification is crucial for minimizing crop damage, improving management practices, and reducing chemical dependence [9].

Deep learning (DL) has significantly advanced the field of plant pathology by enabling accurate, objective, and scalable approaches to disease detection [10]. Convolutional Neural Networks (CNNs) have been key to this progress, effectively learning and extracting relevant features for the classification and identification of plant diseases [11,12]. These architectures support multiple computer vision applications, including image classification, object detection, and instance segmentation [13]. Instance segmentation is especially useful for assessing disease severity and tracking infection progression, as it enables the prediction of pixel-level masks and class labels for each affected region [14,15].

The YOLO (You Only Look Once) architecture was introduced to address the challenge of object detection, which had traditionally been performed in two separate stages. By unifying bounding box regression and classification within a single neural network, YOLO enabled fast end-to-end training, greatly improving detection speed and efficiency [16]. Its early versions (v1–v4) focused solely on single-stage detection with multi-scale feature extraction [17–19], while modern versions (v5–v11) incorporate advanced modules such as C3k2 and C2PSA to enhance feature extraction, spatial attention, and computational efficiency, supporting multiple tasks including instance segmentation, pose estimation, and oriented object detection [20,21].

Transformers architecture was first introduced by Vaswani et al. [22] for Natural Language Processing applications. Dosovitskiy et al. [23] extended this approach to computer vision by proposing the vision transformer (ViT), which demonstrated superior performance over the previously dominant CNNs when trained on sufficiently large datasets. Despite its strong representational capacity, ViT suffers from significant limitations, most notably the quadratic growth in computational complexity with increasing image resolution.

MobileViT was proposed by Mehta and Rastegari [24] as a lightweight hybrid architecture that integrates convolutional layers with Transformer blocks. By leveraging convolutions for local feature extraction and applying self-attention on compact patch representations, MobileViT reduces computational and memory overhead while maintaining competitive accuracy, making it well suited for mobile and edge-device deployments.

Recent advances in smartphone-based imaging and artificial intelligence have enabled citizen science platforms for real-time plant disease and pest monitoring. Christakakis et al. [25] developed a cross-platform mobile application integrated with a cloud-based decision support system (DSS), allowing users to upload field-acquired RGB data for AI-driven pest and disease detection. The system employs deep learning models within a scalable backend infrastructure, enabling continuous data integration and model improvement while supporting post-border surveillance of quarantine and invasive pests. Field validation demonstrated the practicality of the approach, achieving 70.2% $mAP_{50}$ in detecting *Tuta absoluta* in tomato cultivations, with detection confidence levels reaching 87%. These results underscore the potential of mobile, DSS-enabled AI tools to support scalable, data-driven crop protection and sustainable precision agriculture.

Tomato leaf disease detection using maching learning (ML) and DL has been extensively studied, though few focus specifically on *B. cinerea*. Qasrawi et al. [26] evaluated five tomato diseases—including *B. cinerea*—using smartphone images and multiple classifiers, achieving 68.9–70.3% accuracy. Loyani et al. [27] applied Mask R-CNN for semantic and instance segmentation on a custom tomato dataset, reaching a mean average precision at 50% Intersection over Union ($mAP_{50}$) of 85.67%. More recently, Kapetas et al. [28] proposed an early detection pipeline for tomato gray mold using multispectral imaging. The pipeline

combined YOLO-based leaf segmentation, transformer models for feature extraction from leaf masks, and an ensemble classifier for final prediction. YOLOv8 segmentation achieved 81.7% $mAP_{50}$, while the ensemble on transformer-derived descriptors reached 79.41% accuracy and 71.12% F1-score. This approach enabled early detection of latent infections, demonstrating the potential of integrating multispectral imaging with transformer-based DL for early *B. cinerea* detection.

For pepper, early detection of *B. cinerea* has been explored using multispectral imaging and DL. Kapetas et al. [29] employed YOLOv11-Small for leaf segmentation on hyperspectral and RGB data, achieving 86.4% $mAP_{50}$. Leaf descriptors extracted via multiple Transformer models were classified using an ensemble, yielding 87.42% overall accuracy and 81.13% F1-score. Similarly, CNN-based multispectral approaches have achieved over 90% accuracy for early detection of Phytophthora blight in pepper [30], and hyperspectral imaging has been successfully applied for monitoring pest-induced stress in bell pepper, supporting the applicability of spectral imaging combined with ML for disease and stress detection in pepper [31].

Cucumber has some of the most advanced ML/DL pipelines for early *B. cinerea* detection. Giakoumoglou et al. [32] employed multispectral imaging employing U-Net++ with a MobileViT-S encoder, achieving 90.1% overall accuracy, a Dice coefficient of 67.7%, an IoU of 65.6%, and a recall of 80.7%, with early infection detection reaching IoU values between 23.0% and 43.7% at 1–6 days post inoculation (dpi). Building on this, Christakakis et al. [33] optimized a vision transformer-based U-Net++ model with MobileViTV2-125 encoders and incorporated the Cut-and-Paste augmentation method to address class imbalance. This approach increased the representation of infected regions, improving early-stage infection detection. Their model achieved 92% accuracy, a mean Dice coefficient of 79.2%, a mean IoU of 81.6%, and a recall of 88.5%, effectively detecting *B. cinerea* from early, non-visible to visible symptoms. In addition, Wang et al. [34] proposed a two-stage pipeline combining DeepLabV3+ for leaf segmentation and U-Net for infected leaf detection, achieving 93.27% segmentation accuracy and 92.85% disease severity classification accuracy. Mia et al. [35] applied deep learning models to detect multiple cucumber diseases, including downy and powdery mildew, reaching 93.23% accuracy.

Recent studies have also highlighted the effectiveness of MobileViT-based architectures for plant disease detection under field and resource-constrained conditions. Zhang et al. [36] introduced a dual-attention-enhanced MobileViT model for rice disease identification, achieving 99.6% classification accuracy on field-captured data while maintaining low computational cost. Ding and Yang [37] applied MobileViT transfer learning for apple leaf disease classification, reaching 97.3% initial accuracy, which improved to 98.54% after optimized transfer learning strategies. Chowdhury and Ahmed [38] proposed MangoLeafViT, a lightweight vision transformer for mango leaf disease classification, achieving 99.43% accuracy on the MangoLeafBD dataset, with minimal parameter count and reduced FLOPs, demonstrating both high efficiency and state-of-the-art performance suitable for mobile deployment.

Despite considerable progress in deep learning for plant disease detection, most published work targets a single crop at a time. Multispectral imaging paired with vision transformers has produced strong results for *B. cinerea* on tomato, pepper, and cucumber individually, but multispectral cameras remain highly expensive for many growers. A model that relies on ordinary RGB data and covers visible symptoms of *B. cinerea* across all three crops has not been performed.

To tackle these limitations, a multi-crop dataset of RGB data was employed using a two-stage approach of a leaf segmenter and a soft voting ensemble classifier of 13 MobileViT variants to address the detection of infected leaves despite the limited samples. Section 2

details the dataset and training procedure, Section 3 reports the evaluation process, and Sections 4 and 5 discuss limitations and future work.

## 2. Materials and Methods

This section outlines the methodology for multi-crop detection of *B. cinerea* using a two-stage deep learning approach. As illustrated in Figure 1, the proposed solution combines instance segmentation for leaf localization with ensemble classification for disease detection across tomato, pepper, and cucumber crops using RGB imagery. The solution follows a two-stage design. In the first stage, RGB data are processed by a leaf segmentation model to accurately delineate and isolate individual leaf regions from the background, producing pixel-level leaf masks. In the second stage, these masks are fed into multiple leaf-level classification models, each generating a class prediction. The individual predictions are then fused by an ensemble model using a voting-based aggregation strategy to produce a robust final leaf class prediction.

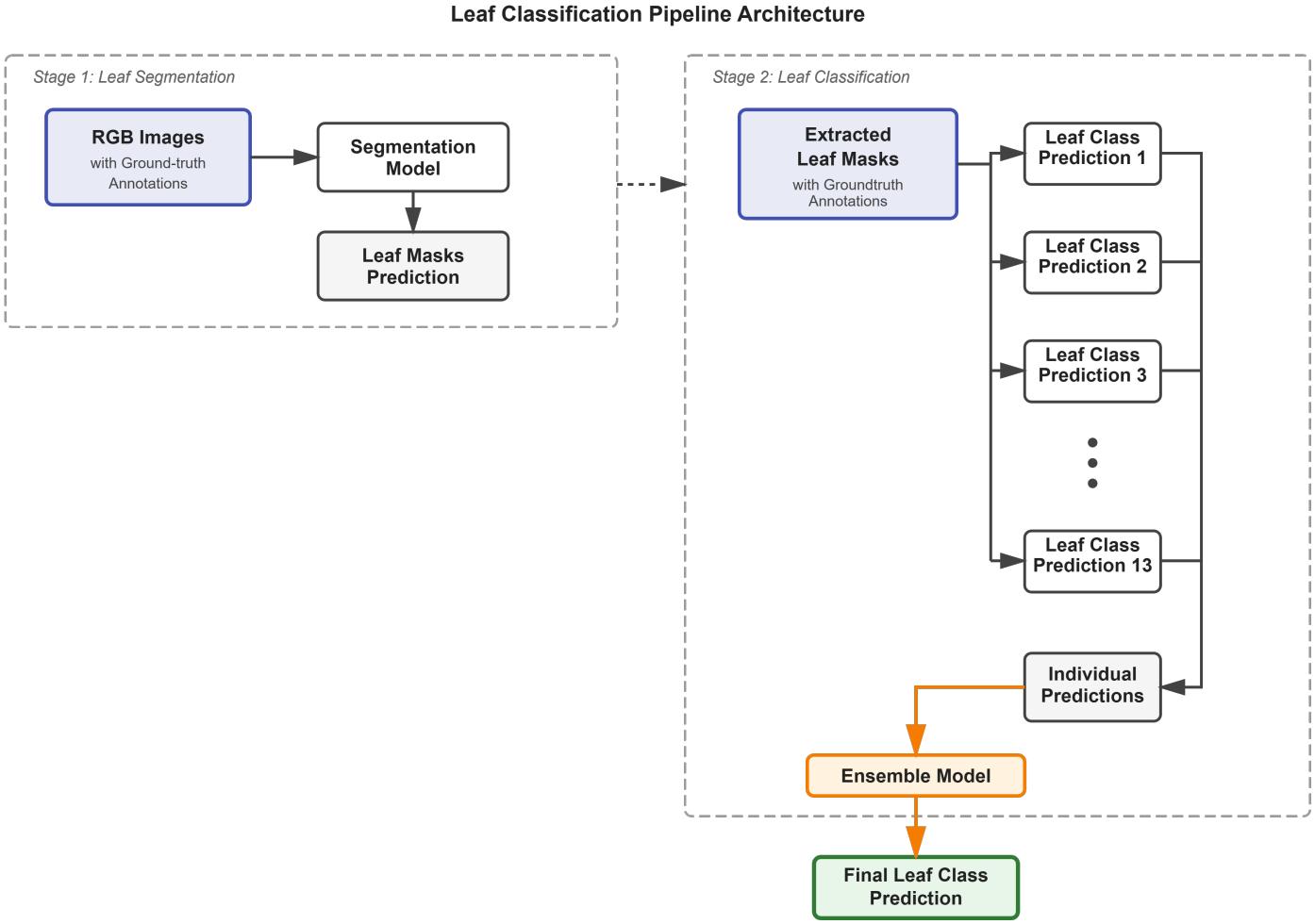



**Figure 1.** Architecture of the two-stage leaf segmentation and classification approach. Color coding in the figure clarifies the role of each component: blue boxes denote input data and ground-truth annotations, white boxes represent models, grey boxes depict model predictions, the orange box highlights the ensemble model, and the green box indicates the leaf class prediction.

### 2.1. Data Collection

*B. cinerea* was isolated from naturally infected cucumber, pepper, and tomato plants and used to artificially inoculate plants. Tomato, pepper, and cucumber plant trials each included six *B. cinerea*-inoculated plants and three mock-inoculated controls. Leaf clas-

sification as "healthy" or "infected" was performed under the guidance of an expert phytopathologist: leaves from mock-inoculated plants or inoculated leaves without visible symptoms were labeled as "healthy," while leaves exhibiting visible infected areas were labeled as "infected." Symptom imaging was carried out using the PhenoCheck camera [39]. Each leaf was annotated using polygonal masks to capture its full area. Leaves were categorized based on infection. In total, 274 tomato leaf captures, 667 pepper leaf captures, and 1061 cucumber leaf captures were annotated using Roboflow [40]. Figure 2 shows representative annotated samples from the three crops.

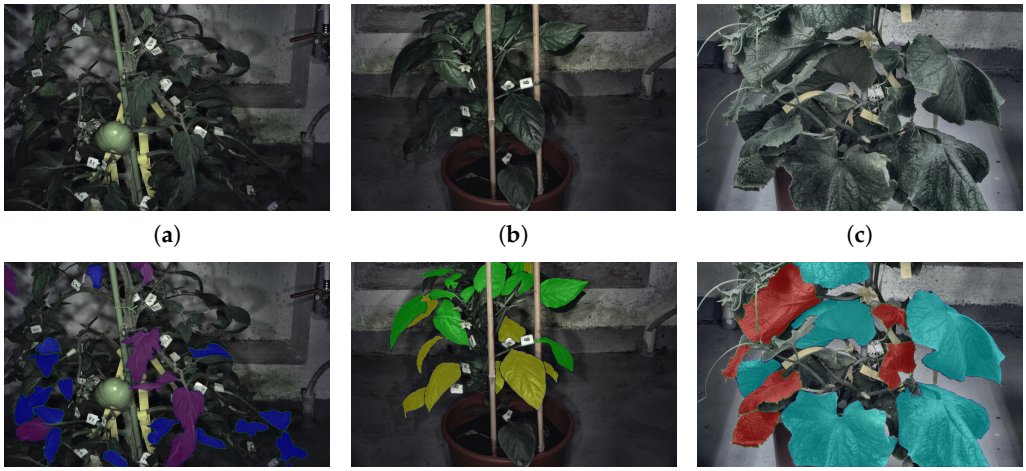


**Figure 2.** Original RGB images (top row) and ground-truth leaf masks with annotations (bottom row) for (**a**,**d**) tomato, (**b**,**e**) pepper, and (**c**,**f**) cucumber. Color coding: tomato healthy (blue), tomato infected (purple), pepper healthy (green), pepper infected (yellow), cucumber healthy (red), cucumber infected (cyan).

### *2.2. Dataset Preprocessing and Augmentation*

#### 2.2.1. Segmentation Dataset Preprocessing

For instance, for segmentation, the dataset comprised plant RGB images annotated at the pixel level across three crop types: tomato, pepper, and cucumber. Each leaf instance was delineated using polygon masks, distinguishing between healthy and infected tissue. Instances of very small leaves (<0.2%), whose area was trivial relative to image dimensions, were filtered out due to insufficient discriminative information for reliable segmentation.

Captures were divided into training and validation sets on a per-day basis, using an iterative approach in order to ensure proper representation of each class on each day in the final dataset. The resulting split contains 18,667 training instances (73.56%) and 6710 validation instances (26.44%), as shown in Table 1.

**Table 1.** Dataset class distribution statistics. Percentages represent the proportion of total instances.

| Crop | Condition | Set | Count | Proportion of Total |
|---|---|---|---|---|
| Tomato | Healthy | Train | 3982 | 15.69% |
| Tomato | Healthy | Valid | 1581 | 6.23% |
| Tomato | Infected | Train | 803 | 3.16% |
| Tomato | Infected | Valid | 105 | 0.41% |
| Pepper | Healthy | Train | 7548 | 29.75% |
| Pepper | Healthy | Valid | 2229 | 8.78% |
| Pepper | Infected | Train | 149 | 0.59% |
| Pepper | Infected | Valid | 19 | 0.07% |

**Table 1.** *Cont.*

| Crop | Condition | Set | Count | Proportion of Total |
|---|---|---|---|---|
| Cucumber | Healthy | Train | 4262 | 16.79% |
| Cucumber | Healthy | Valid | 2052 | 8.09% |
| Cucumber | Infected | Train | 1923 | 7.58% |
| Cucumber | Infected | Valid | 724 | 2.85% |
| Total | | Train | 18,667 | 73.56% |
| Total | | Valid | 6710 | 26.44% |

#### 2.2.2. Classification Dataset Preprocessing and Augmentation

For classification, a leaf-level dataset was created from the segmentation dataset by extracting individual leaf instances using ground truth segmentation masks and labels with transparent backgrounds. All leaf images were resized to 224 × 224 pixels to match the input dimensions of MobileViT architectures.

Dataset diversity was enhanced with data augmentation on the training set. The augmentation pipeline included random horizontal and vertical flips, and color transformations with brightness, contrast, and saturation adjustments, each with a variation factor of 0.2. These choices follow common practice in agricultural image classification [11,12] and are meant to approximate natural variation in leaf orientation and lighting.

### *2.3. Model Training*

#### 2.3.1. Segmentation Model

For leaf localization, instance segmentation was performed using YOLOv11 variants, specifically YOLOv11L-seg (Large) and YOLOv11X-seg (Extra-Large), both configured for single-class leaf segmentation. Training was conducted over up to 50 epochs with early stopping to prevent overfitting. All models processed input images at a resolution of 1024 × 1024 pixels. During the first 30 epochs, mosaic augmentation was applied to enhance data diversity; it was subsequently disabled to allow the model to focus on cleaner feature representations in later training stages. The stochastic gradient descent (SGD) optimizer was employed with an initial learning rate of 0.01, reduced gradually to 0.001 via a cosine scheduler, and supplemented by a weight decay of 0.00537. To identify optimal training parameters, batch sizes of 2, 4, and 8 were systematically evaluated.

#### 2.3.2. Classification Model

A total of 13 MobileViT model variants were employed in this study to comprehensively assess the performance of hybrid CNN-Transformer architectures, where convolutional layers extract local features and lightweight transformer blocks capture global dependencies. The naming convention (XXS, XS, S, 0.5, 0.75, 1.0, 1.25, 1.75, 2.0) reflects a systematic scaling of network dimensions. The complete list of employed models and their pre-training strategies is detailed in Table 2.

To mitigate the class imbalance documented in Table 1, inverse-frequency weighting was applied to the classification loss. Classes with fewer training samples receive proportionally larger loss penalties, which reduces the model bias to the dominant healthy classes, and forces the model to focus on the unrepresented classes. The architectural choice of MobileViT for classification is also motivated in part by the nature of the task: unlike purely convolutional backbones, MobileViT's hybrid design pairs local convolutional features with lightweight Transformer blocks that capture long-range spatial context across the leafs.

As shown in Table 2, the evaluation includes three standard MobileViT v1 variants (XXS, XS, S) and ten MobileViT v2 variants with different width scaling factors (0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0). The v2 models are further divided into those with standard

ImageNet-1K pre-training and three larger variants (`mobilevitv2_150`, `mobilevitv2_175`, and `mobilevitv2_200`) that underwent additional pre-training on the larger ImageNet-22K dataset before fine-tuning on ImageNet-1K. This design allows for a direct analysis of the impact of both architectural scaling (width multiplier) and pre-training data scale on final classification performance.

**Table 2.** Employed MobileViT model variants and their pre-training strategy.

| Model Family | Model Identifier | Pre-Training Strategy |
|---|---|---|
| MobileViT v1 | mobilevit_xxs.cvnets_in1k | ImageNet-1K only |
| | mobilevit_xs.cvnets_in1k | ImageNet-1K only |
| | mobilevit_s.cvnets_in1k | ImageNet-1K only |
| MobileViT v2 (Standard Pre-training) | mobilevitv2_050.cvnets_in1k | ImageNet-1K only |
| | mobilevitv2_075.cvnets_in1k | ImageNet-1K only |
| | mobilevitv2_100.cvnets_in1k | ImageNet-1K only |
| | mobilevitv2_125.cvnets_in1k | ImageNet-1K only |
| | mobilevitv2_150.cvnets_in1k | ImageNet-1K only |
| | mobilevitv2_175.cvnets_in1k | ImageNet-1K only |
| | mobilevitv2_200.cvnets_in1k | ImageNet-1K only |
| MobileViT v2 (Extended Pre-training) | mobilevitv2_150.cvnets_in22k_ft_in1k | ImageNet-22K + ImageNet-1K fine-tuning |
| | mobilevitv2_175.cvnets_in22k_ft_in1k | ImageNet-22K + ImageNet-1K fine-tuning |
| | mobilevitv2_200.cvnets_in22k_ft_in1k | ImageNet-22K + ImageNet-1K fine-tuning |

Classification performance was further improved through the application of ensemble learning strategies. Three distinct voting methods were implemented: soft voting, in which predicted class probabilities are averaged across models; hard voting, where the final class label is determined by majority voting; and weighted hard voting, in which individual model votes are weighted proportionally to their corresponding F1-scores. These ensemble techniques leverage the complementary strengths of different model variants to achieve more robust and accurate predictions than any single model could provide alone.

### *2.4. Model Evaluation*

The evaluation of segmentation and classification models utilized distinct metrics appropriate to their respective tasks at the pixel and image levels. Segmentation performance was measured by mean average precision at 50% Intersection over Union ($mAP_{50}$), mean Intersection over Union (mIoU), and mean Dice Similarity Coefficient (mDSC). Classification performance was assessed using precision (P), recall (R), and the F1-score (F1). The corresponding mathematical definitions are provided in Table 3.

**Table 3.** Performance metrics for segmentation and classification.

| Metric | Formula | Description |
|---|---|---|
| Segmentation Metrics | | |
| $mAP_{50}$ | $\frac{1}{N}\sum AP_i^{50}$ | Mean average precision at an IoU threshold of 0.5. |
| mIoU | $\frac{1}{N}\sum \frac{TP_i}{TP_i+FP_i+FN_i}$ | Mean Intersection over Union, quantifying spatial overlap between predicted and ground-truth masks. |
| mDSC | $\frac{1}{N}\sum \frac{2TP_i}{2TP_i+FP_i+FN_i}$ | Mean Dice Similarity Coefficient, similar to mIoU but assigning higher weight to true positives. |

**Table 3.** *Cont.*

| Metric | Formula | Description |
|---|---|---|
| Classification Metrics | | |
| $P$ | $\frac{TP}{TP+FP}$ | Proportion of correctly predicted positive samples (true positives (TPs)) among all predicted positives (TPs and false positives (FPs)). |
| $R$ | $\frac{TP}{TP+FN}$ | Proportion of correctly predicted positive samples (TPs) among all actual positives (TPs and false negatives (FNs)). |
| $F_1$ | $\frac{2PR}{P+R}$ | Harmonic mean of precision and recall, balancing false positives and false negatives. |

## 3. Results

### *3.1. Performance Evaluation of the Segmentation Model*

Table 4 summarizes the segmentation performance of different YOLO11 configurations across batch sizes. The YOLO11l-seg model with a batch size of 4 produced the best results overall, reaching a Mask $mAP_{50}$ of 87.3%, an mIoU of 76.8%, and an mDSC of 86.0%. The larger YOLO11x-seg model, despite its greater parameter capacity, scored below the lighter YOLO11l-seg variants at every batch size tested. With the limited training data available in this study, the extra capacity of the X-seg model appears to work against it: the model fits noise in the training set instead of learning patterns that transfer to unseen data.

**Table 4.** YOLO11 single-class segmentation model performance comparison.

| Model Size | Batch Size | Mask $mAP_{50}$ (%) | mIoU (%) | mDSC (%) |
|---|---|---|---|---|
| yolo11l-seg | 2 | 86.90 | 76.67 | 85.6 |
| **yolo11l-seg** | **4** | **87.30** | **76.77** | **86.00** |
| yolo11l-seg | 8 | 86.70 | 76.90 | 86.00 |
| yolo11x-seg | 2 | 86.80 | 76.57 | 85.60 |
| yolo11x-seg | 4 | 83.30 | 75.90 | 85.60 |
| yolo11x-seg | 8 | 86.60 | 76.30 | 85.60 |

Bold indicates the best-performing model among batch and model size.

### *3.2. Performance Evaluation of the Classification Model*

#### 3.2.1. Individual Models

Table 5 details the classification performance of individual MobileViT variants. MobileViTv2-0.5 achieved the highest overall F1 of 80.17%. The relationship between model size and accuracy is non-monotonic: MobileViTv2-0.5 outperforms both the smaller v1 variants, which may lack the capacity to capture enough discriminative features, and the larger v2 models, which show signs of overfitting on a dataset of this size. Beyond a certain width, the added parameters pick up training noise instead of generalizable patterns, a trend consistent with the same effect observed on the segmentation side.

Looking at the infected classes specifically, pepper infected reached the highest F1 across all models, peaking at 93.33% with MobileViTv2-0.5, thanks to the sharp necrotic margins that make *B. cinerea* lesions on pepper easy to distinguish. Tomato infected followed at up to 82.68%, where symptoms are visible but less contrasted against the leaf surface. Cucumber infected proved hardest, topping out at 69.00% with MobileViTv2-0.5. The diffuse spread of the pathogen along cucumber veins and the high variability in leaf texture and pose make infected and healthy cucumber leaves difficult to tell apart at $224 \times 224$ resolution.

**Table 5.** MobileViT model performance comparisons.

| Model | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|
| MobileViT-XXS | 75.91 | 74.26 | 72.97 |
| MobileViT-XS | 76.14 | 75.94 | 74.91 |
| MobileViT-S | 72.82 | 74.68 | 68.43 |
| **MobileViTv2-0.5** | **82.67** | **80.93** | **80.17** |
| MobileViTv2-0.75 | 79.57 | 78.22 | 77.15 |
| MobileViTv2-1.0 | 76.10 | 80.21 | 76.91 |
| MobileViTv2-1.25 | 78.00 | 80.46 | 78.17 |
| MobileViTv2-1.5 | 78.08 | 73.82 | 74.88 |
| MobileViTv2-1.5 (in22k_ft_in1k) | 84.48 | 75.69 | 76.70 |
| MobileViTv2-1.75 | 72.81 | 74.94 | 73.21 |
| MobileViTv2-1.75 (in22k_ft_in1k) | 74.75 | 75.75 | 72.80 |
| MobileViTv2-2.0 | 76.64 | 78.62 | 75.55 |
| MobileViTv2-2.0 (in22k_ft_in1k) | 73.26 | 68.76 | 69.66 |

Bold indicates the best-performing model among the MobileViT variants pretrained on ImageNet-1k.

Figure 3 presents the confusion matrix for the YOLO11l-seg model's single-class leaf segmentation task. The model shows strong localization with 5863 TPs. It misclassified 1874 background regions as leaves (FPs), indicating a tendency to over-segment by including stems. Conversely, 846 leaf instances were missed (FNs), likely due to occlusion, lighting, or subtle edges.

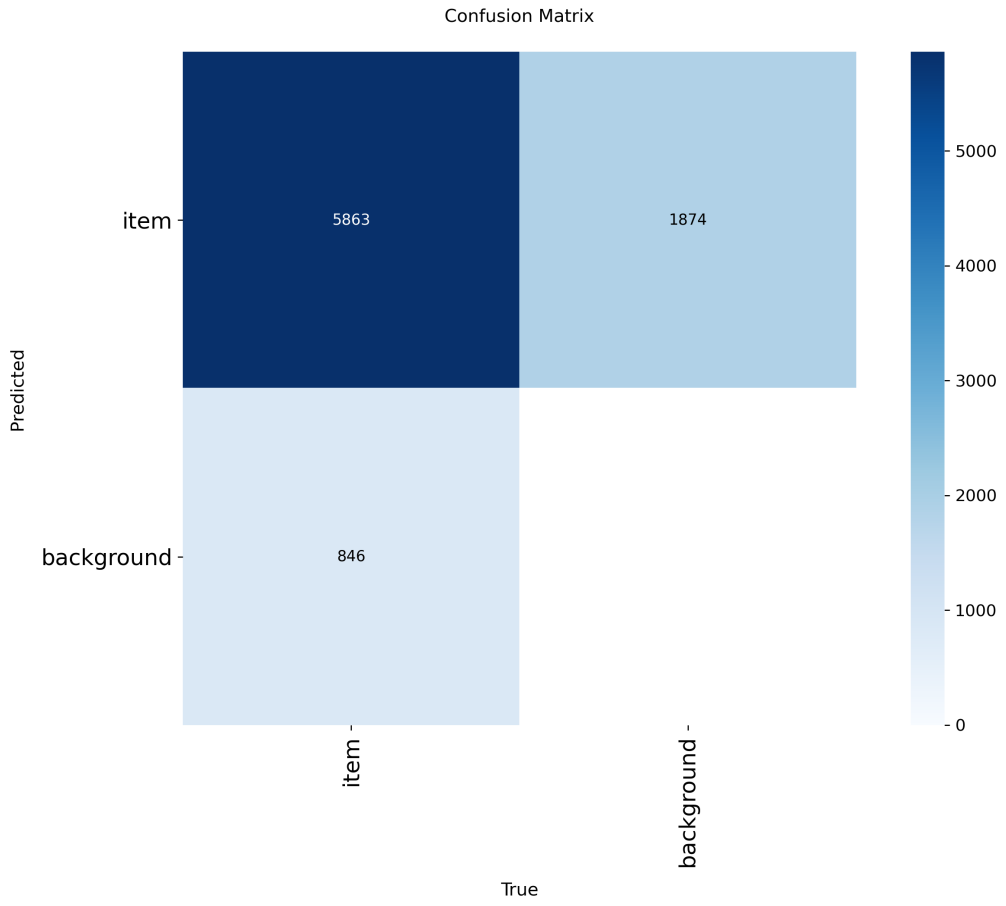



**Figure 3.** Confusion matrix for YOLO11l-seg instance segmentation showing leaf (item) versus background classification.

### 3.2.2. Ensemble Model

The ensemble-based classification results, detailed in Table 6, demonstrate that model ensemble yields a consistent performance improvement over the best individual model. Among the three strategies, soft voting emerged as the most effective, achieving the highest overall F1-score of 84.05%. This represents a gain of +3.90% over the top-performing individual model, MobileViT-0.5 (F1-score: 80.17%).

Class-level analysis reveals that soft voting provides the most balanced improvements. For the challenging cucumber infected class, it improved the F1-score by +1.50% compared to the best individual model, and also delivered notable gains for cucumber healthy (+1.80%) and tomato infected (+1.20%). This indicates that probabilistic fusion stabilizes predictions for classes with subtle or ambiguous visual symptoms.

In contrast, hard voting and weighted hard voting achieved smaller overall improvements (+2.90 and +2.80%, respectively) and showed mixed effects at the class level. Specifically, both strategies reduced the F1-score for the pepper infected class by −3.30%. At the same time, it improved the cucumber infected class—the worst-performing class in the best individual model—by +2.70%. Overall, soft voting remains the optimal ensemble strategy, providing greater robustness and more consistent per-class performance, which is essential for reliable multi-crop disease detection.

Figure 4 presents the confusion matrix for the soft voting ensemble classifier. The diagonal shows strong performance for healthy classes. The primary confusion lies within crops, between healthy and infected states: infected cucumber has 81 FPs, infected pepper has 4 FPs, and infected tomato has 55 FPs.

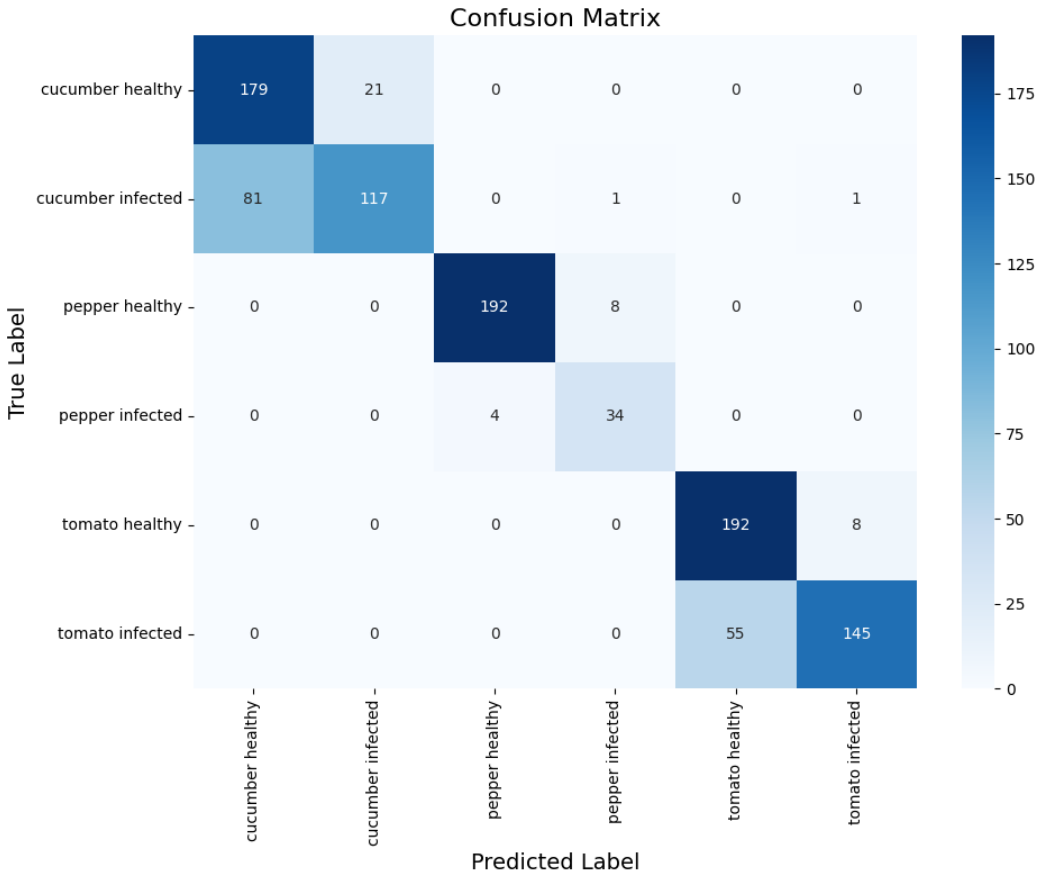



**Figure 4.** Confusion matrix for the soft voting ensemble across classes.

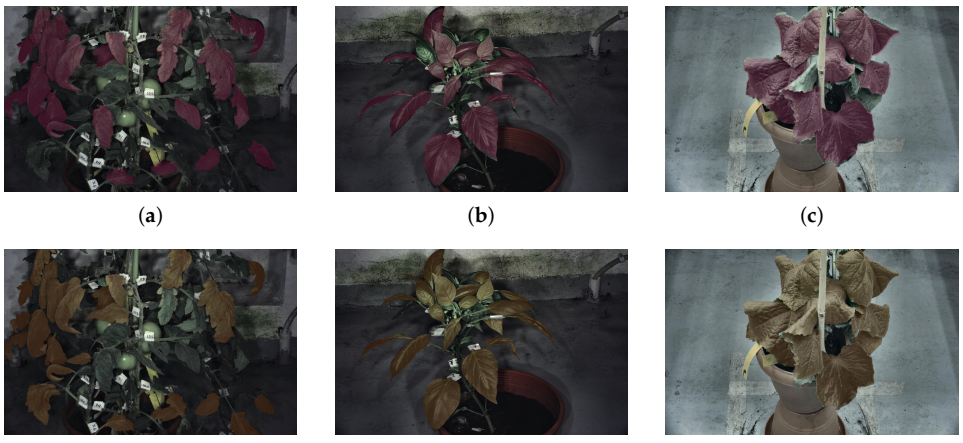
**Table 6.** Ensemble method performances compared to best individual model ($F1_{Difference} = F1_{Ensemble} - F1_{Best\ Individual}$).

| Method/Class | Precision (%) | Recall (%) | F1-Score (%) | $F1_{Diff.}$ (%) |
|---|---|---|---|---|
| **Soft Voting—Overall** | **85.10** | **84.77** | **84.05** | **3.90** |
| **cucumber healthy** | **69.80** | **89.00** | **78.24** | **1.80** |
| **cucumber infected** | **84.62** | **60.50** | **70.55** | **1.50** |
| **pepper healthy** | **98.48** | **97.00** | **97.73** | **1.00** |
| **pepper infected** | **85.37** | **92.11** | **88.61** | **1.20** |
| **tomato healthy** | **78.69** | **96.00** | **86.49** | **1.50** |
| **tomato infected** | **93.67** | **74.00** | **82.68** | **1.20** |
| Hard Voting—Overall | 83.05 | 84.00 | 83.03 | 2.90 |
| cucumber healthy | 72.44 | 81.50 | 76.71 | 0.30 |
| cucumber infected | 78.61 | 68.00 | 72.92 | 2.40 |
| pepper healthy | 97.94 | 95.00 | 96.45 | −0.30 |
| pepper infected | 77.27 | 89.47 | 82.93 | −3.30 |
| tomato healthy | 78.93 | 95.50 | 86.43 | 1.50 |
| tomato infected | 93.13 | 74.50 | 82.78 | 1.30 |
| Weighted Hard Voting—Overall | 83.13 | 84.08 | 83.12 | 2.80 |
| cucumber healthy | 72.77 | 81.50 | 76.89 | 0.50 |
| cucumber infected | 78.74 | 68.50 | 73.26 | 2.70 |
| pepper healthy | 97.94 | 95.00 | 96.45 | −0.30 |
| pepper infected | 77.27 | 89.47 | 82.93 | −3.30 |
| tomato healthy | 78.93 | 95.50 | 86.43 | 1.50 |
| tomato infected | 93.13 | 74.50 | 82.78 | 1.30 |

Bold indicates the best-performing ensemble method (Soft Voting) and its corresponding per-class results.

#### 3.2.3. Qualitative Results

Figure 5 shows a qualitative comparison between the predicted leaf segmentation masks produced by the YOLO11-seg model and the corresponding ground truth annotations. The predictions illustrate the model's strong capability to accurately delineate leaf boundaries across all crop types. Mask predictions closely align with the ground truth in terms of shape, coverage, and spatial consistency, even in cases of partial occlusion. Minor deviations occur mainly at leaf edges or in highly occluded regions, where thin structures or severe overlap may lead to misdetected leaves; however, these cases are limited and do not significantly affect the overall correspondence between predictions and ground truth.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

**Figure 5.** Visualization of predicted and ground truth leaf segmentation masks for (**a**,**d**) tomato, (**b**,**e**) pepper, and (**c**,**f**) cucumber. The top row (**a**–**c**) displays segmentation predictions produced by the trained YOLO11l-seg model; the bottom row (**d**–**f**) shows the corresponding ground truth annotations.

Figure 6 presents a qualitative visualization of the leaf-level classification results from the soft voting ensemble, enabling a direct comparison between predicted class labels

and ground truth annotations for tomato, pepper, and cucumber samples. The ensemble demonstrates robust performance in distinguishing both crop species and health status, with predictions closely matching ground truth labels and reliably identifying healthy and infected leaves. Most misclassifications occur in leaves with subtle visual differences between healthy and infected regions; nonetheless, the ensemble effectively reduces prediction noise and enhances label consistency compared to individual models.

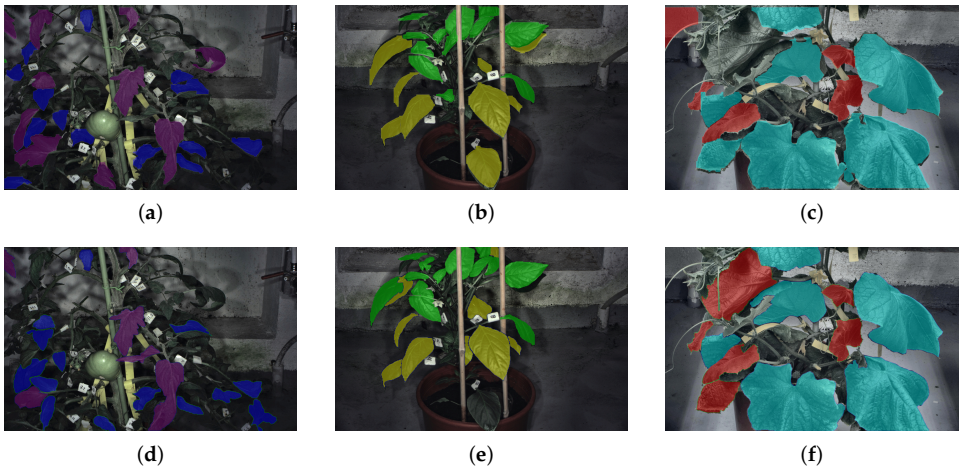



**Figure 6.** Predicted and ground truth instance segmentation masks with healthy/infected leaf classifications for (**a**,**d**) tomato, (**b**,**e**) pepper, and (**c**,**f**) cucumber. The top row (**a**–**c**) shows predictions from the YOLO11l-seg model and soft-voting ensemble; the bottom row (**d**–**f**) shows the corresponding ground truth annotations. Color coding: tomato healthy (blue), tomato infected (purple), pepper healthy (green), pepper infected (yellow), cucumber healthy (red), cucumber infected (cyan).

Figure 7 illustrates the per-crop performance differences. For pepper, infected leaves display distinctive necrotic lesions with well-defined margins, leaving little room for ambiguity; the few false negatives are cases where lesions are small. Tomato occupies a middle ground: false negatives tend to be leaves with only faint or localized symptoms that push the model close to its decision boundary. Cucumber represents the hardest case. Correctly classified infected cucumber leaves show lighter, water-soaked patches, but the same heterogeneous texture, curled edges, and dark venation patterns appear in both correctly and incorrectly classified groups.

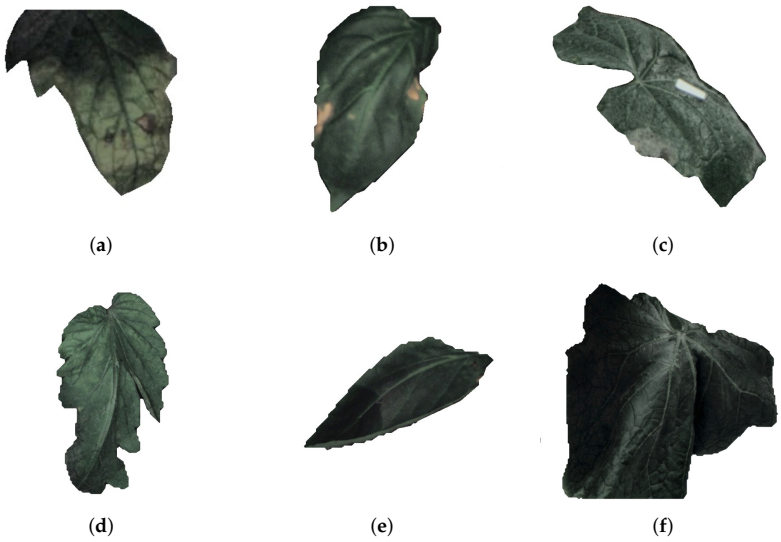



**Figure 7.** Representative correctly classified (**top row**) and misclassified (**bottom row**) infected leaf samples from the soft-voting ensemble for (**a**,**d**) tomato, (**b**,**e**) pepper, and (**c**,**f**) cucumber.

### *3.3. Computational Performance*

Training was performed on an NVIDIA A100 GPU. Early stopping patience was set to 5 epochs, and therefore training times vary across classifiers. Also, inference latencies are averaged over the entire validation set. Table 7 reports training and inference times for every model in the pipeline.

**Table 7.** Computational profile of all evaluated models. Segmentation latency is reported at 1024 × 1024. Classification metrics are per 224 × 224 leaf patch. All measurements on an NVIDIA A100 GPU.

| Model | Train Time (min) | Inference (ms) |
|---|---|---|
| Segmentation—YOLO11L-seg | | |
| 1024 × 1024 | 28.28 | 25.1 |
| Classification—Individual MobileViT models (per 224 × 224 patch) | | |
| MobileViT-XXS | 8.0 | 14.9 |
| MobileViT-XS | 6.1 | 16.4 |
| MobileViT-S | 6.1 | 14.6 |
| MobileViTv2-0.5 | 5.5 | 10.0 |
| MobileViTv2-0.75 | 3.7 | 10.0 |
| MobileViTv2-1.0 | 4.9 | 10.3 |
| MobileViTv2-1.25 | 7.4 | 10.1 |
| MobileViTv2-1.5 | 3.6 | 10.6 |
| MobileViTv2-1.5 (22k) | 4.9 | 10.3 |
| MobileViTv2-1.75 | 11.0 | 10.7 |
| MobileViTv2-1.75 (22k) | 5.5 | 10.2 |
| MobileViTv2-2.0 | 12.3 | 10.7 |
| MobileViTv2-2.0 (22k) | 7.4 | 10.7 |
| 13-model soft-voting ensemble (per patch, sequential) | | |
| All 13 models | — | 149.5 |
| Summary | | |
| Segmentation (train/inference) | 28.28 | 25.1 |
| Classification ensemble (train/inference) | 86.4 | 149.5 |
| End-to-end per leaf (inference) | — | 174.6 |

Therefore, the pipelines require on average 25 ms of inference time to acquire the YOLO instance segmentation predictions for a single image and another 150 ms of inference time to compute the ensemble voting output for all leaves, bringing the total inference time per image to approximately 175 ms.

## 4. Discussion

Table 8 sets this study's results beside published *B. cinerea* detection studies. Direct comparison is complicated by differences in datasets, imaging modality, number of target crops, and evaluation protocols, but the numbers are still informative.

On the segmentation side, this study's YOLO11L-seg reaches 87.3% $mAP_{50}$ and 86.0% mDSC for single-class leaf detection across all three crops using RGB input only. An important caveat is that the two cucumber *B. cinerea* segmentation studies solve a different problem from this study. Christakakis et al. [33] used U-Net++ with a MobileViTv2-125 encoder and the Cut-and-Paste augmentation method for three-class (healthy, invisible infection and visible infection) disease-region segmentation on cucumber leaves, reaching 79.2% mDSC, 81.6% mIoU, and 92% pixel accuracy on RGB and multi-spectral data. Giakoumoglou et al. [32] applied U-Net++ with MobileViT-S to the same task, and obtained 67.7% mDSC, 65.6% IoU, and 90.1% pixel accuracy, also on RGB and multi-spectral data.

Therefore, direct metric comparison (even for the exact same metric) between these two studies and this study, although informative and useful, is not exactly representative of model performance.

**Table 8.** Comparison with related *B. cinerea* detection studies. Seg. = segmentation; Cls. = classification.

| Study | Imaging | Crops | Seg. metric | Cls. Metric |
|---|---|---|---|---|
| Giakoumoglou et al. [32] | Multi-sp. | Cucumber | mDSC 67.7%, IoU 65.6% | Pixel acc. 90.1% |
| Christakakis et al. [33] | Multi-sp. | Cucumber | mDSC 79.2%, mIoU 81.6% | Pixel acc. 92% |
| Kapetas et al. [28] | Multi-sp. | Tomato | $mAP_{50}$ 81.7% | Acc. 79.4%, F1 71.1% |
| Kapetas et al. [29] | Multi-sp. + RGB | Pepper | $mAP_{50}$ 86.4% | Acc. 87.4%, F1 81.1% |
| **This study** | **RGB** | **Tomato, Pepper, Cucumber** | **$mAP_{50}$ 87.3%** | **F1 84.05%** |

Bold indicates the proposed method (this study).

Among studies that also perform leaf-level segmentation, [28] obtained 81.7% $mAP_{50}$ with YOLOv8, while their subsequent study (Kapetas et al. [29]) achieved 86.4% $mAP_{50}$ with YOLOv11-Small on pepper using a combination of multi-spectral plus RGB data. Both these studies focus on single-crop data, whereas this study covers three crops and exceeds with 87.3% mAP50 and solely RGB data usage.

On the classification side, only two (Kapetas et al. studies [28,29]) of the four comparative studies offer per-leaf classification like this study. Christakakis et al. and Giakoumoglou et al. both operate at the pixel level on cucumber and do not report leaf-level accuracy or F1-score. Kapetas et al. [29] achieved 87.42% accuracy and 81.13% F1-score on pepper using a four-Transformer ensemble (Swin-L, ViT-L, VOLO, XCIT-L) with LSTM classification on multi-spectral data plus five vegetation indices; per-class F1-scores were 85.25% (healthy), 66.67% (botrytis-invisible), and 78.26% (botrytis-visible). Their earlier tomato work [28] reached 79.41% accuracy and 71.12% F1-score with a similar four-Transformer ensemble (MaxViT-L, ViT-B, VOLO, XCIT-L) and LSTM classification; per-class F1-score was 68.33% (healthy), 63.71% (invisible), and 77.47% (visible). This study's soft-voting ensemble of 13 MobileViT variants scores 84.05% F1-score across three crops from RGB only. On a per-crop basis, this study's F1-score (86.61%) for the pepper-infected class exceeds [29] best pepper per-class F1-score (85.25%). Similarly, this study's F1-score (82.68%) for the tomato-infected class sits well above their tomato pipeline's overall F1-score of 71.12%. Cucumber is the weakest performing crop of this study, with an F1-score of 70.55%. No prior *B. cinerea* study reports a cucumber-specific classification result, so a direct baseline is not available. The broad pattern is consistent: prior work reaches high accuracy on one crop with specialized multi-spectral cameras, while the pipeline handles three crops with an ordinary camera at competitive F1-scores.

The class imbalance documented in Table 1 affects each crop differently. For the Pepper data, the class with the fewer samples (149 training instances) still achieved 88.61% F1-score. *B. cinerea* symptoms on pepper produce sharp necrotic margins that are easy to separate from healthy tissue, and the inverse-frequency loss weighting (Section 2) kept the model from ignoring this minority class. For the cucumber data, on the other hand, with 1923 samples, the model only achieved a 70.55% F1-score. The bottleneck is not data volume but symptom appearance. More balanced data through oversampling or synthetic generation could enable the model to reach even higher performance levels and this remains a topic for future study. The deeper issue, though, remains visual ambiguity between symptomatic and asymptomatic samples.

The uneven image counts reflect differences in plant material availability during the trials. However, the disparity is partially offset by differences in leaf complexity across crops. Tomato and pepper leaves share a similar morphology, relatively flat lamina, and analogous venation patterns, which means fewer captures are needed to capture their visual variability.

Their classification scores are correspondingly close (F1 82.68% and 88.61% for infected leaves, respectively). Cucumber presents a harder problem regardless of dataset size: its leaves are considerably wider, often curl at the edges and overlap one another, occluding symptomatic tissue. These morphological and pathological differences, rather than the smaller tomato training pool alone, are the main drivers of the per-crop performance gaps.

The augmentation choices applied to the classifier (inverse-frequency loss weighting, random flips, color jitter, and resized crops) follow standard practice for image classification under class imbalance and were selected based on established guidelines in the literature. The augmentation were applied uniformly to all classification models, while a deep dive into the exact performance benefits from each augment fell out of the scope of the current work.

A practical concern for deployment is error propagation. The segmentation model produces 1874 false positives (mostly stems), and in an end-to-end run these non-leaf patches would be forwarded to the classifier. However, given that the classifier was trained exclusively on leaf patches, it has no learned representation for stems or background and is likely to return unreliable predictions for such inputs. Filtering out these cases, for instance through a secondary leaf-vs.-non-leaf check or a score threshold on the segmentation output, was not implemented here or targeted by this study, but remains a topic for future work.

Further experiments that were performed but are not reported in detail in this study include YOLO multi-class detection, predicting all six crop-disease labels in one pass. This approach yielded poor performance on minority classes and was therefore disregarded as an option. Splitting the pipeline into a balanced single-class segmentation task and a separately weighted classification task proved better.

All our imagery data utilized in this study were captured with a customized PhenoCheck imaging setup. Field conditions (variable lighting, cluttered backgrounds, different phone cameras, and wind-blown leaves) will cause a domain shift. Closing that gap calls for progressive fine-tuning on field-collected data, stronger augmentation (random lighting, background noise, motion blur), confidence thresholds that flag uncertain cases for human review, and possibly an incremental-learning loop so the model improves as users confirm or correct its predictions.

The pipeline structure is appropriate for integration into a decision support system in a straightforward way. The final approach would include the farmer capturing images of leaves with a smartphone and uploading the data to the cloud; there, the two-stage model proposed in this study would perform inference runs, and the output per-leaf results (health status, confidence score, annotated overlay) would appear on screen, with the system flagging infected plants for targeted treatment. The citizen-science framework described by Christakakis et al. [25] already provides a detailed methodological approach for such a deployment.

## 5. Conclusions

This paper presented a two-stage pipeline that detects visible *B. cinerea* infection on tomato, pepper, and cucumber leaves from RGB data. YOLO11L-seg segments individual leaves at 87.3% $mAP_{50}$, and a 13-model MobileViT soft-voting ensemble classifies each leaf as healthy or infected at 84.05% overall F1-score. Per-crop infected-leaf F1-score is 88.61% for pepper, 82.68% for tomato, and 70.55% for cucumber. The full training pipeline completes in under two hours on a single A100 GPU, and end-to-end inference takes about 175 ms per leaf.

Compared with prior *B. cinerea* detection studies, which target one crop at a time and rely on multi-spectral cameras, this study's system covers three crops with RGB data at a

competitive F1-score. The trade-off is that only visible symptoms are detected; latent or early-stage infections that lack surface cues fall outside the capability of RGB-only input.

Analysis of the results indicated some clear limitations. Model performance for the cucumber-infected class remains the weakest, with B. cinerea symptoms that have not fully progressed, resembling normal leaf visual characteristics. All data were collected with a customized PhenoCheck multispectral camera under controlled conditions, so field performance under variable lighting, cluttered backgrounds, and different cameras still needs validation.

An independent test set was not held out, as the minority classes' sample data were already scarce (149 infected-pepper samples, 105 infected-tomato samples). This scarcity does not support a reliable three-way split. (train, validation, split). This study utilized image augmentations with uniform parameters for the classifiers during training according to current dedicated research. Thus, a dedicated performance comparative evaluation for different augmentation parameters was deemed not necessary.

Analysis of this study's results indicates further steps for future work, specifically for the cucumber data, to address the relatively lower performance levels for that crop. Such further steps could include further data capture, experimentation with (higher input resolutions to the models, attention-guided cropping, or multi-scale patch analysis.)

Progressive fine-tuning on field-collected data would test the pipeline's robustness outside controlled settings. Adding a lightweight leaf-vs.-non-leaf gate between the two stages could filter segmentation artifacts. Finally, the modular design of the pipeline allows either stage to be replaced independently; for instance, incorporating multi-spectral input into the classification stage could extend coverage to latent infections without redesigning the segmentation model.

**Author Contributions:** Conceptualization: E.M.P.; methodology: C.R.M., D.K., E.M.P.; software: C.R.M.; validation: D.K.; formal analysis: C.R.M.; investigation: C.R.M. and D.K.; resources: E.M.P.; data curation: C.R.M. and D.K.; writing—original draft preparation: C.R.M.; writing—review and editing: D.K. and E.M.P.; visualization: C.R.M.; supervision: E.M.P.; project administration: E.M.P.; funding acquisition: E.M.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data will be made available on request.

**Acknowledgments:** This work is carried out within AISE project no. 26702, under the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union—Next Generation EU (Implementation body: HFRI) within the 3rd Call for H.F.R.I. Research Projects to support Faculty Members and Researchers. We thank Benaki Phytopathological Institute and iKnowHow for their invaluable technical support in molecular experiments and assistance in data collection.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. FAO. World Food and Agriculture Statistical Yearbook 2022. Available online: https://openknowledge.fao.org/server/api/core/bitstreams/fba4ef43-422c-4d73-886e-3016ff47df52/content (accessed on 20 February 2026).
2. More Peppers Than Ever Produced in 2022. Available online: https://www.hortidaily.com/article/9590006/more-peppers-than-ever-produced-in-2022/ (accessed on 20 February 2026).
3. Poveda, J.; Barquero, M.; González-Andrés, F. Insight into the microbiological control strategies against Botrytis cinerea using systemic plant resistance activation. *Agronomy* **2020**, *10*, 1822. [CrossRef]

4. Hua, L.; Yong, C.; Zhang, Z.; Li, B.; Qin, G.; Tian, S. Pathogenic Mechanisms and Control Strategies of *Botrytis Cinerea* Causing Post-Harvest Decay Fruits Vegetables. *Food Qual. Saf.* **2018**, *2*, 111–119. [CrossRef]
5. Latorre, B.A.; Elfar, K.; Ferrada, E.E. Gray Mold Caused by *Botrytis cinerea* Limits Grape Production in Chile. *Cienc. Investig. Agrar.* **2015**, *42*, 316–330.
6. Romanazzi, G.; Smilanick, J.L.; Feliziani, E.; Droby, S. Integrated Management of Postharvest Gray Mold on Fruit Crops. *Postharvest Biol. Technol.* **2016**, *113*, 69–76. [CrossRef]
7. Leroux, P.; Fritz, R.; Debieu, D.; Albertini, C.; Lanen, C.; Bach, J.; Gredt, M.; Chapeland, F. Mechanisms of Resistance to Fungicides in Field Strains of *Botrytis cinerea*. *Pest Manag. Sci.* **2002**, *58*, 876–888. [CrossRef]
8. Rosslenbroich, H.-J.; Stuebler, D. *Botrytis cinerea*—History of chemical control and novel fungicides for its management. *Crop Prot.* **2000**, *19*, 557–561. [CrossRef]
9. Bilkiss, M.; Shiddiky, M.J.A.; Ford, R. Advanced Diagnostic Approaches for Necrotrophic Fungal Pathogens of Temperate Legumes with a Focus on *Botrytis* spp. *Front. Microbiol.* **2019**, *10*, 1889. [CrossRef]
10. Wäldchen, J.; Mäder, P. Machine Learning for Image-Based Species Identification. *Methods Ecol. Evol.* **2018**, *9*, 2216–2225. [CrossRef]
11. Toda, Y.; Okura, F. How Convolutional Neural Networks Diagnose Plant Disease. *Plant Phenom.* **2019**, *2019*, 9237136. [CrossRef] [PubMed]
12. Barburiceanu, S.; Meza, S.; Orza, B.; Malutan, R.; Terebes, R. Convolutional Neural Networks for Texture Feature Extraction: Applications to Leaf Disease Classification in Precision Agriculture. *IEEE Access* **2021**, *9*, 160085–160103. [CrossRef]
13. Geng, Q.; Zhou, Z.; Cao, X. Survey of Recent Progress in Semantic Image Segmentation with CNNs. *Sci. China Inf. Sci.* **2018**, *61*, 051101. [CrossRef]
14. Hafiz, A.M.; Bhat, G.M. A Survey on Instance Segmentation: State of the Art. *Int. J. Multimed. Inf. Retr.* **2020**, *9*, 171–189. [CrossRef]
15. Charisis, C.; Argyropoulos, D. Deep Learning-Based Instance Segmentation Architectures in Agriculture: A Review of the Scopes and Challenges. *Smart Agric. Technol.* **2024**, *8*, 100448. [CrossRef]
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2016; pp. 779–788. [CrossRef]
17. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2017; pp. 7263–7271. [CrossRef]
18. Farhadi, A.; Redmon, J. YOLOv3: An Incremental Improvement. *Comput. Vis. Pattern Recognit.* **2018**, *1804*, 1–6.
19. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
20. Sapkota, R.; Qureshi, R.; Calero, M.F.; Hussain, M.; Badjugar, C.; Nepal, U.; Poulose, A.; Zeno, P.; Vaddevolu, U.B.P.; Yan, H.; et al. YOLO11 to Its Genesis: A Decadal and Comprehensive Review of the You Only Look Once Series. *arXiv* **2024**, arXiv:2406.19407v5.
21. Ultralytics. Ultralytics YOLO Documentation. 2025. Available online: https://docs.ultralytics.com (accessed on 20 February 2026).
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
24. Mehta, S.; Rastegari, M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
25. Christakakis, P.; Papadopoulou, G.; Mikos, G.; Kalogiannidis, N.; Ioannidis, D.; Tzovaras, D.; Pechlivani, E.M. Smartphone-Based Citizen Science Tool for Plant Disease and Insect Pest Detection Using Artificial Intelligence. *Technologies* **2024**, *12*, 101. [CrossRef]
26. Qasrawi, R.; Amro, M.; Zaghal, R.; Sawafteh, M.; Polo, S.V. Machine Learning Techniques for Tomato Plant Diseases Clustering, Prediction and Classification. In *2021 International Conference on Promising Electronic Technologies (ICPET)*; IEEE: New York, NY, USA, 2021; pp. 40–45. [CrossRef]
27. Loyani, L.K.; Bradshaw, K.; Machuve, D. Segmentation of Tuta Absoluta's damage on tomato plants: A computer vision approach. *Appl. Artif. Intell.* **2021**, *35*, 1107–1127. [CrossRef]
28. Kapetas, D.; Kalogeropoulou, E.; Christakakis, P.; Klaridopoulos, C.; Pechlivani, E.M. Multi-spectral image transformer descriptor classification combined with molecular tools for early detection of tomato grey mould. *Smart Agric. Technol.* **2024**, *9*, 100580. [CrossRef]
29. Kapetas, D.; Kalogeropoulou, E.; Christakakis, P.; Klaridopoulos, C.; Pechlivani, E.M. Comparative Evaluation of AI-Based Multi-Spectral Imaging and PCR-Based Assays for Early Detection of Botrytis cinerea Infection on Pepper Plants. *Agriculture* **2025**, *15*, 164. [CrossRef]

30. Duan, Z.; Li, H.; Li, C.; Zhang, J.; Zhang, D.; Fan, X.; Chen, X. A CNN Model Early Detect. Pepper Phytophthora Blight Using Multispectral Imaging, Integrating Spectral and Textural Information. *Plant Methods* **2024**, *20*, 115. [CrossRef]
31. Krüger, M.; Zemanek, T.; Wuttke, D.; Dinkel, M.; Serfling, A.; Böckmann, E. Hyperspectral imaging for pest symptom detection in bell pepper. *Plant Methods* **2024**, *20*, 156. [CrossRef]
32. Giakoumoglou, N.; Kalogeropoulou, E.; Klaridopoulos, C.; Pechlivani, E.M.; Christakakis, P.; Markellou, E.; Frangakis, N.; Tzovaras, D. Early detection of Botrytis cinerea symptoms using deep learning multi-spectral image segmentation. *Smart Agric. Technol.* **2024**, *8*, 100481. [CrossRef]
33. Christakakis, P.; Giakoumoglou, N.; Kapetas, D.; Tzovaras, D.; Pechlivani, E.-M. Vision Transformers in Optimization of AI-Based Early Detection of Botrytis cinerea. *AI* **2024**, *5*, 1301–1323. [CrossRef]
34. Wang, C.; Du, P.; Wu, H.; Li, J.; Zhao, C.; Zhu, H. A cucumber leaf disease severity classification method based on the fusion of DeepLabV3+ and U-Net. *Comput. Electron. Agric.* **2021**, *189*, 106373. [CrossRef]
35. Mia, M.J.; Maria, S.K.; Taki, S.S.; Biswas, A.A. Cucumber disease recognition using machine learning and transfer learning. *Bull. Electr. Eng. Informat.* **2021**, *10*, 3432–3443. [CrossRef]
36. Zhang, M.; Lin, Z.; Tang, S.; Lin, C.; Zhang, L.; Dong, W.; Zhong, N. Dual-attention-enhanced MobileViT network: A lightweight model for rice disease identification in field-captured images. *Agriculture* **2025**, *15*, 571. [CrossRef]
37. Ding, Y.; Yang, W. Classification of apple leaf diseases based on MobileViT transfer learning. In *International Conference on Image Processing and Artificial Intelligence (ICIPAI 2024)*; SPIE: Bellingham, WA, USA, 2024. [CrossRef]
38. Chowdhury, R.H.; Ahmed, S. MangoLeafViT: Leveraging lightweight vision transformer with runtime augmentation for efficient mango leaf disease classification. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*; IEEE: New York, NY, USA, 2024. [CrossRef]
39. Spectricon. PhenoCheck—Handheld Spectral Phenotyper. 2025. Available online: https://spectricon.com/products-and-markets/phenocheck/ (accessed on 20 February 2026).
40. Roboflow. Roboflow: Image Annotation and Dataset Management Platform. 2025. Available online: https://roboflow.com (accessed on 20 February 2026).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.