International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(7s) 2026 pp. 1284-1300
https://doi.org/10.70917/ijcisim-2026-3196
Article

Self-Supervised Multimodal Vision Transformer Framework for Robust Tomato Leaf Disease Detection under Field Conditions

Annagiri Amarnath[1], S. K. Mahaboob Basha[2]

[1]Department of Computer Science, Bharatiya Engineering Science & Technology Innovation University, Gorantla, Andhra Pradesh, India.
Email: annagiri.amar@gmail.com
[2]Department of Computer Science and Engineering, Sree Dattha Institute of Engineering and Science, Hyderabad, Telangana, India.
Email: email2mahaboob@gmail.com

Abstract: Prompt and precise identification of tomato leaf diseases are essential towards enhancing crop production and minimizing losses in precision farming. But current image-based disease detection algorithms are mostly based on fully supervised convolutional neural networks that are trained on controlled datasets and this restricts their resilience and extrapolation in real-field problems with complicated backgrounds, variances of illumination, and scarce labeled samples. In order to overcome these difficulties, this paper introduces a single deep learning model, which incorporates self-supervised representation learning, Vision Transformer (ViT) architecture, attention-based multimodal feature fusion, domain adaptation, and few-shot learning in terms of real-field tomato leaf disease detection.

The offered solution uses self-supervised pretraining to obtain discriminative visual representations using unlabeled field images, thus very minimally reliant on big labeled datasets. Adaptively fused multimodal features such as the visual RGB cues, texture as well as vegetation color indices are combined to improve disease discrimination in harsh environmental conditions through an attention mechanism. Domain adaptation is added to reduce the problems of dataset bias, and few-shot learning can be used to classify the categories of diseases with high precision with a small number of labeled samples.

The framework is tested on tomato leaf images in the real field and compared to the known CNN and transformer-based baselines in accordance with the accuracy, precision, recall, F1-score, and confusion matrix analysis. The experimental findings show that the suggested method performs better as far as robustness and generalization are concerned especially when the data are limited and the tasks are cross-domain. This is also the first research to unite self-supervised learning of Vision Transformers, multimodal fusion, and domain adaptation to detect tomato leaf disease in the real field in a systematic manner, making it well-suited to be deployed in real-life agriculture applications.

Keywords: unsupervised learning; Multimodal fusions of features; Vision Transformer; Plant diseases; Tomato leaf diseases; Domain adaptation; Few-shot learning; Precision agriculture

1. INTRODUCTION

Tomato (Solanumlycopersicum) is a vegetable crop that is commonly grown and has a high economic value in most countries of the world. It is highly susceptible to a large number of

foliar diseases that impact its productivity and quality severely, and unless they are detected at an early stage, the losses of the yields can be significant. Existing disease detection techniques are based on visual observations of the agricultural specialists, which is tedious, subjective, and unfeasible (when it comes to mass monitoring). This has led to the development of automated disease detection systems which are based on images and is an area of research important in precision agriculture.

Deep learning methods, especially convolutional neural networks (CNNs), have shown good results in plant disease classification in the past few years. There have been numerous studies that reported high accuracy with supervised CNN models that were trained on benchmark datasets like PlantVillage. Although these are reported successes, most of the current methods are designed and tested in controlled conditions of imaging at smooth backgrounds and with large numbers of labelled samples. These models tend to exhibit severe performance losses when applied into a real-field setting because of fluctuation with illumination, occlusion, background clutter, leaf orientation and inter-class similarity.

A second inherent weakness of existing ways of doing things is that they rely too heavily on large volumes of labeled data. Obtaining good annotations to use in agricultural images is also expensive and needs expert knowledge, especially those of rare or emerging plant diseases. This weakness renders fully supervised learning strategies inapplicable to scalable and practical agricultural implementations. Moreover, CNN-based networks are biased towards local texture features, and they usually do not represent global contextual correlations which are essential to differentiate similar disease symptoms in the field.

Recently Vision Transformers (ViTs) have received interest as an alternative to CNNs because they can model long-range dependencies, as they apply self-attention mechanisms. Nevertheless, ViT models are usually trained on big labeled datasets and this means that they cannot be applied directly to agricultural spheres. Self-supervised learning (SSL) provides a potentially effective response to this problem since it provides models with the ability to learn meaningful learnable representations of features using unlabeled data. SSL can greatly minimize the need to label the field images as well as enhance the generalization capabilities by taking advantage of the abundance of unlabeled field images.

Besides the difficulty with representation learning, real-field disease detection systems require the efficient use of a combination of visual cues. The symptoms of the disease are characterized by color changes, but also by variations in texture and patterns in their structure. The single-modality learning methods often cannot reflect such complexity. When multimodal features learning is paired with attention processes, models are able to dynamically focus on the most informative features, and thus become much more resilient to environmental changes. Furthermore, domain adaptation methods are necessary in order to minimize the bias in the dataset and in order to make sure that the learned representations are practical in new acquisition contexts.

This paper was inspired by these difficulties and offers a cohesive deep learning model of the real-field detection of tomato leaf diseases, which combines self-supervised Vision Transformer pretraining, attention-based multi-modal feature fusion, domain adaptation, and few-shot learning in a single framework. Unlike the previous research which highlights the use of performance on controlled datasets, this paper concentrates on realistic analysis of deployment viability through use of real-field images only.

Contributions and Novelty

As its major contributions, one can list the following:

1. An integrated framework is suggested, which combines self-supervised learning of Vision Transformer, multimodal feature fusion, domain adaptation and few-shot classification to detect tomato leaf diseases.
2. The given approach is tested only on actual-field tomato leaf images, which is relevant to the real issues that are not disregarded in the majority of the existing literature.
3. Attention-based multimodal fusion approach is applied to merge adaptively the RGB, texture, and vegetation color features, which will increase the discrimination in a complex environmental condition.
4. The framework is shown to be more robust and generalization performances in the case of small-scale labeled data, thus it is applicable to scalable precision agriculture.

2. Related Work

Detection of tomato leaf diseases with computer vision and machine learning methods has been a commonly studied topic. The current research might be generally divided into supervised CNN-based, transformer-based, self-supervised learning, multimodal feature learning, and domain adaptation with few-shot learning. This part is a critical review of these directions and their inability to be applied to the real field.

2.1 Supervised CNN-Based Tomato Leaf Disease Detection

Thomato leaf disease classification Early deep learning models used mainly supervised convolutional neural networks (CNNs) including AlexNet, VGGNet, GoogLeNet, and ResNet. These models were found to have great classification accuracy when trained and tested on standard dataset like plantvillage with scores well beyond 95. Yet, these findings are mostly dependent on the fact that such data sets are controlled, in that they have homogeneous backgrounds, the same lighting, and solitary leaves. Accordingly, CNN-based models trained on such data conditions exhibit high levels of performance decrease in the case of real-field images with cluttered backgrounds, shadows, occlusions, and variable lighting.

Lightweight CNNs like MobileNet, ShuffleNet and EfficientNetwere developed to overcome the limitations of computational efficiency and deployment. Despite the fact that EfficientNet-based models were able to achieve better accuracy-efficiency trade-off, they were, in fact, still highly reliant on massive amounts of labeled data and could not be as robust to domain shift. In addition, the majority of CNN-based approaches are purely based on RGB images and, therefore, are vulnerable to the background noise and cannot be sufficiently expressive to identify the subtle signs of the disease under field conditions.

2.2 Transformer-Based Vision Models for Plant Disease Detection.

Vision Transformers (ViTs) are relatively new substitutes to CNNs because they are capable of representing long-range dependencies and global contextual information. Some papers

implemented supervised ViT architectures on the tasks of plant and tomato leaf disease classification, and their performance was equal or slightly higher than that of CNN-based models. Although ViTs exhibit better contextual reasoning, current methods are similarly based on supervised training on labeled RGB images with the same annotation dependency and generalization weaknesses as CNNs.

More than that, supervised ViT models are both data-intensive and susceptible to overfitting in case of training on reasonably small size agricultural datasets. Their representational ability is not fully harnessed without large scale labeled pretraining, and their resistance to real-field behavior is also uneven. Consequently, other systems of tomato disease detection based on supervised transformer have yet to prove to be reliable in a non-constrained agricultural setting.

2.3 Self-Supervised Learning in Agricultural Vision.

Self-supervised learning (SSL) has received interest as a way to acquire meaningful visual representations using unlabeled data, without incurring the expensive cost of annotation. Some of the techniques that have demonstrated good transferability to different computer vision tasks include SimCLR, MoCo, DINO, DINOv2, and Masked Autoencoders (MAE). SSL has been investigated in agriculture to classify crops, phenotype plants, and also to analyze vegetation in general.

Nevertheless, there are few studies where self-supervised foundation models are used to spot tomato leaf disease. Available literature tends to either deploy small-scale SS systems or test the performance in controlled environments, but they do not consider variability in the real-field or changing domains. Furthermore, there is a paucity of data that has been presented on the combination of multimodal feature integration or few-shot learning techniques with the use of SSL in agricultural disease detection, and its full capabilities have not been fully realized.

2.4 Multimodal Feature Learning for Disease Recognition.

The symptoms of diseases in tomato leaves are visualized because of a mixture of the visual indicators, such as lesion texture, discoloring patterns, and areas of irregularity. Although RGB images can record world appearance, it is often hard to isolate finer physiological stress measures, even though it may be required. In order to deal with this, some of the studies considered texture characterization like GLCM and LBP and vegetation color indices like Excess Green and normalized color ratios.

Although they have proven to be useful, the majority of multimodal methods use naive feature concatenation or shallow fusion algorithms, which implicitly assume some equal weight between modalities. These strategies do not represent cross-modal interactions and tend to perform poorly in complicated field environments. Multimodal fusion, in particular, attention-based, has been shown to be very effective in medical imaging and remote sensing, but has not been used extensively in the detection of tomato leaf diseases.

2.5 Domain Adaptation and Few-Shot Learning.

Domain shift between the laboratory data and actual field images has remained a challenge in the agricultural vision systems. Controlled training models are frequently poor in learning in the presence of unknown conditions of acquisition. Because of their capability to reduce the impact of this problem, domain adaptation methods such as feature-level alignment and adversarial learning have been demonstrated to be effective but are not systematically applied to tomato disease detection systems.

Moreover, tomato diseases are rare, and they have acute data imbalance, which causes biased classifiers on the supervised environment. Few-shot learning can provide a means of generalizing over limited labeled samples; nevertheless, little use is made of it in practice in the detection of tomato leaf disease. The vast majority of the current systems focus on the overall accuracy on dominant classes and not on the robustness when the quantity of available data is small.

2.6 Research Gap and Motivation

In the analysis above, some of the drawbacks of current literature can be identified:

The majority of tomato leaf disease detection approaches are based on fully supervised learning, and they need large volumes of labeled data.

CNN- and supervised ViT-based models are not very robust in real field under domain shift. The multimodal information is either not used or combined in ineffective strategies.

Both domain adaptation and few-shot learning are seldom combined into a single detection system.

It is little explored how self-supervised Vision Transformer models can be used to detect tomato disease in the real world.

This paper is inspired by these shortcomings and aims to present a self-supervised multimodal Vision Transformer architecture incorporating attention-driven feature fusion, domain adaptation, and few-shot learning to achieve effective tomato leaf disease recognition in a real-field setting.

3. Methodology

This part outlines the suggested self-supervised multimodal model of detecting tomato leaf disease in the field environment. The design directly tackles three consistent failures of the current systems namely: dependence on large labeled data, poor sensitivity to environmental variation and incapability to operate in rare disease classes. It is a modular methodology where any of the components can be validated and substituted independently of the other components disrupting the whole pipeline.

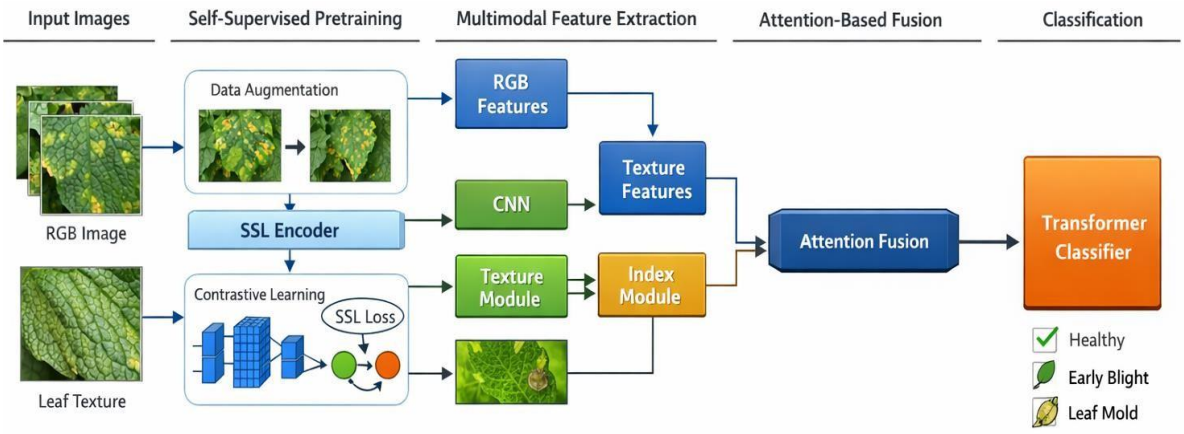

Figure 1 – Overall System Architecture

3.1 Problem Definition

Since the images of tomato leaf were taken under uncontrolled field conditions, the task is to classify all the tomato leaf images correctly to fall within C disease metrics, including healthy leaves. Mathematically, given an input image I the task is to learn a function:

$$f(I) \rightarrow y, \quad y \in \{1, 2, \ldots, C\}$$

where f( ) should be able to generalize to different illumination, backgrounds, severity of disease and domain changes between lab and field and where f () should depend minimum on labeled training data.

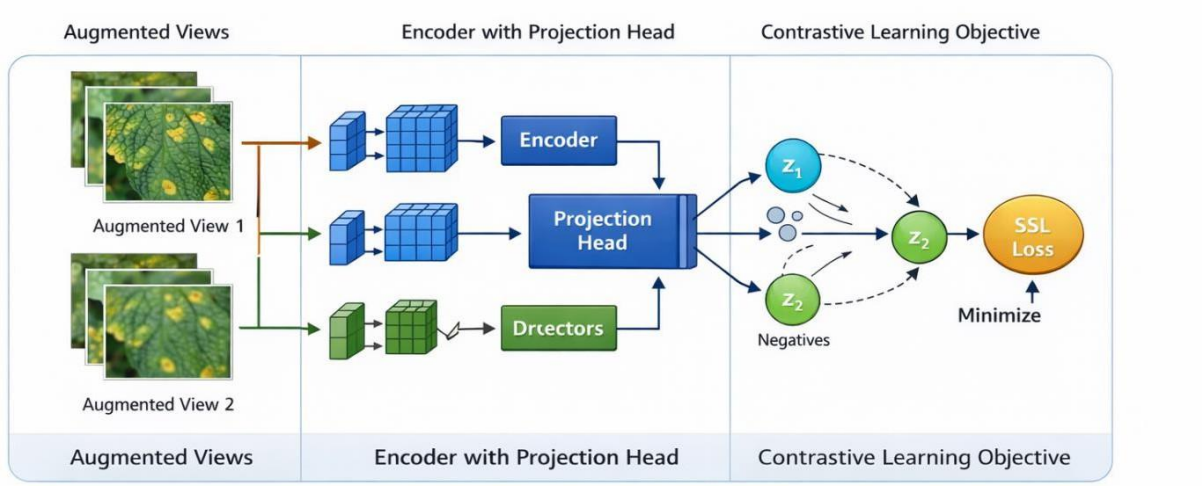

Figure 2. Self-Supervised Learning Workflow

3.2 Overall Framework Overview

The suggested framework is composed of four large parts:

1. Representation learning Self-supervised Self-supervised representation learning with a Vision Transformer backbone.
2. Multimodal extraction of feature based on RGB, texture and vegetation color indices.
3. Multimodal feature fusion based on attention.
4. Few-shot learning and domain adaptation of robust classification.

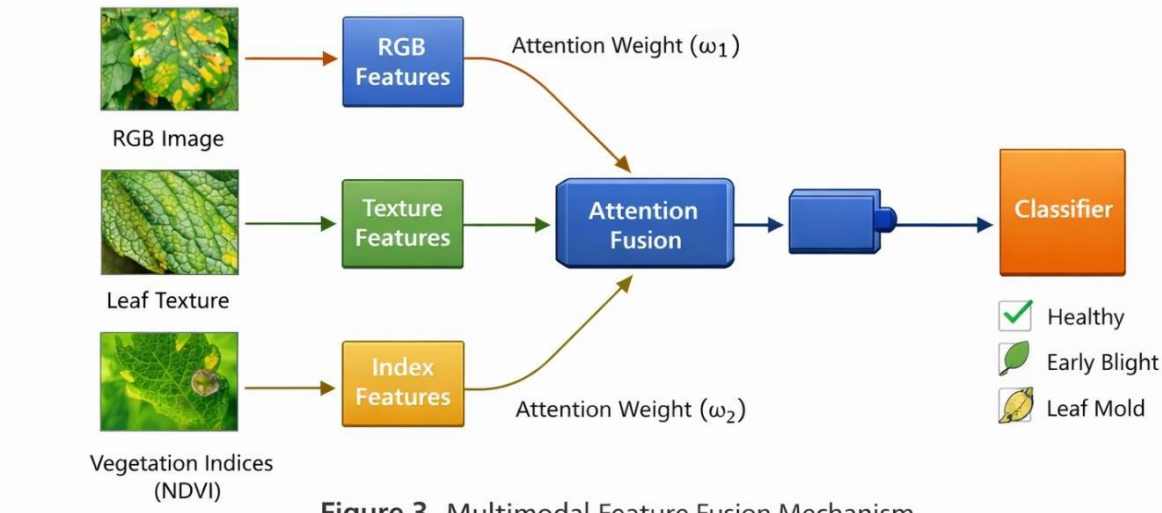

Figure 3. Multimodal Feature Fusion Mechanism

This entire pipeline will aim to initially learn powerful visual representations using unlabeled data and then use them to supervised disease classification using a small number of labeled samples.

3.3 Self-Supervised Pretraining Using Vision Transformers

To reduce the dependency on large labelled samples and increase the generalization performance in real field cases, a Vision Transformer (ViT) backbone is trained with the assistance of self-supervised learning. Self-supervised learning unlike supervised pretraining gives the model an opportunity to learn not just invariant but also transferable representations with just unlabeled tomato leaf images.

In this article, ViT-B/16 network is the backbone network used. The model is trained with two complementary self-supervised methods, namely, DINOv2 and Masked Autoencoders (MAE). In the course of DINOv2 pretraining, a student-teacher structure is employed, in which the teacher network parameters are updated based on an exponential moving average of the student parameters. The aim is to match the distribution of outputs of the student and teacher networks with various image augmentations, which promotes the resistance to lighting, viewpoint, and background changes typical of field images.

In the case of MAE pretraining, the patches of input images have a high masking ratio, where the model is required to infer missing visual data based on surrounding signals. With this approach, the ViT is able to record the global leaf structure and disease spatial patterns. The MAE goal reduces the reconstruction error of the original and reconstruction patch image.

Pretraining Self-supervised pretrainingis carried out on a large set of untagged tomatoes leaf images to be captured in a variety of field conditions. The ViT backbone is then pretrained and then used as a feature extractor to another downstream disease classification task. This two phase training plan will provide a representation that is not biased to particular disease names and can also withstand different acquisition settings.

3.4 Multimodal Feature Extraction

The appearance of the tomato leaves is not only the symptom of the disease by the RGB appearance. Three modalities are extracted in order to obtain complementary diagnostic cues:

3.4.1 *RGB Visual Features*

RGB images give the shape of it globally, the distribution of lesions, and the color. The features are directly extracted using the trained ViT backbone.

3.4.2 *Texture Features*

Texture descriptors represent lesion roughness, spot density and spatial irregularity which is usually essential in differentiating diseases with similar appearance. The representations of textures are computed by applying to them statistical texture operators and coded into feature vectors that are aligned with ViTembeddings.

3.4.3 *Vegetation Color Indices*

Vegetation indices bring out physiological stress and chlorophyll degradation. The indices (common: Excess Green (ExG), GreenRed Difference and normalized color ratios) are calculated to highlight color changes which are caused by the disease but which are not visually prominent in the RGB space.

The modalities are complementary and each one of the modalities captures different information. The use of RGB as a treatment alone has been known to be an oversimplification in the analysis of plant diseases and this is specifically avoided in this work.

3.5 Attention-Based Multimodal Feature Fusion.

Naive feature concatenation is based on the assumption that modalities are equally important, which is falsely evident under the conditions of real-fields. In response to this, attention based fusion mechanism is utilized.

Frgb,Ftex and Fci represent RGB, texture and color index feature embeddings respectively. A module of attention learns modality specific weights:

$$F_{fused} = \alpha_{rgb} F_{rgb} + \alpha_{tex} F_{tex} + \alpha_{ci} F_{ci}$$

where a dynamic learning takes place of attention coefficient aalpha.

The mechanism helps the model to focus on the use of texture cues to indicate the disease with stronger patterns of lesions and the color index to indicate the disease with chlorosis or discoloration. The fusion strategy enhances the discriminative power whilst removing redundant or noisy features.

3.6 Attention-Based Multimodal Feature Fusion.

The symptoms of tomato leaf disease are heterogeneous in nature; they are manifested in lesion texture, color changes and spatial anomalies. In order to encode these complementary features, it is divided into three modalities (RGB images), texture features, and vegetation color features.

Let Frgb[?]Rd ,Ftex[?]Rd and Fci[?]Rd represent feature representations of the RGB, texture, and color index images, respectively. Rather than blindly concatenating features, a dynamic weighting of the contribution of each modality is used, which is referred to as an attention-based fusion mechanism.

Attention weights are calculated to be as:

$$\alpha_i = \frac{\exp(W_i F_i)}{\sum_{j \in \{rgb,tex,ci\}} \exp(W_j F_j)}$$

Wi is a projection matrix to be learned and αi is aligned with alpha i is the normalized weight of the importance of modality i.

Fused feature representation Ffusedis gotten as:

$$F_{fused} = \alpha_{rgb} F_{rgb} + \alpha_{tex} F_{tex} + \alpha_{ci} F_{ci}$$

By formulating it, the network can accentuate patterns of texture of lesion-based diseases, and hue indices of the symptoms of chlorosis, and dishearten the noisy or less informative modalities. Attention-based fusion mechanism enhances discriminative ability and stability in the field with complex conditions.

3.7 Domain Adaptation Strategy

The ability to recognize the difference in the domain shift between the laboratory and the field images is one of the primary weaknesses of the present tomato leaf disease detection systems as it leads to a reduction in the performance of the systems. To mitigate this difficulty, the specified framework consists of domain adaptation on the feature level.

Where the source (laboratory) and target (real-field) domains are denoted by Ds and Dt respectively. Although the differences between the visual distributions of these domains are enormous, the disease semantics of such domains are identical. The merged feature representations are balanced by using a domain discrepancy loss on the merged representations.

The difference between the source and target feature distributions are reduced with the assistance of Maximum Mean Discrepancy (MMD) metric:

$$\mathcal{L}_{DA} = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(F^s_{fused}) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(F^t_{fused}) \right\|^2$$

In which the ph([?]) is a kernel mapping function and the nsnsns and ntntnt are the number of samples in the source and target domains respectively.

Through joint minimization of LDA with classification loss, domain-invariant representations are learned by the model, and they are also stable to new acquisition settings. This approach is important in minimizing the decreased performance when models that have been trained using lab data are used in actual field environments.

3.8 Few-Shot Learning for Rare Disease Classes.

Rare leaf diseases of tomatoes are usually not adequately represented in labeled datasets, and this result in biased classifiers in the fully supervised setting. A few-shot learning strategy is incorporated into the classification stage in case of data-scarce conditions in order to enhance its recognition.

The pretrainedViT backbone is a powerful feature extractor that allows effective generalization of limited labeled samples when used in the self-supervised and unsupervised version. A prototype-based classification strategy is used, with a prototype-vector of a disease class being the average of support features of that disease class.

The class prototype Pf of a given class c cwhose K labeled samples are denoted by C:

$$P_c = \frac{1}{K} \sum_{k=1}^{K} F_{fused}^{(k)}$$

In inference, query sample is categorized in terms of its closeness to prototypes of the class using cosine similarity. This would provide a stable classification even when there are a few labeled samples of a rare disease category.

The few-shot learning component does not need complete retraining of the underlying network, which allows it to be computationally efficient and be used in the field of agriculture in reality.

3.9 Optimization Objective

The general training goal is formulated as a weighted loss of classification and domain adaptation:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{DA}$$

Lcls Lambda is used to denote categorical cross-entropy loss and λ is used to moderate domain adaptation.

4. Description of data set and experimental set up.

In this section, the datasets, data preparation process, experimental setup and evaluation measures that were adopted to determine the performance of the proposed self-supervised multimodal tomato leaf disease detection framework are described. These details should be clearly specified so that they can be reproduced and compared to the currently available methods.

The dataset was collected and described using the following method.

4.1 Dataset Collection andDescriptionThe following method was used to collect and describe the dataset.

The evaluation of the experiment is done on a curated tomato leaf disease image dataset which includes samples that were taken under the controlled and real-field conditions. The data has five categories, which are Healthy, Early Blight, Late Blight, Leaf Mold, and Septoria. Real-field images had been directly taken of tomato plants that were planted in open agricultural conditions, with natural differences in the level of illumination, background clutter, leaf orientation, and occlusion, as well as the severity of the disease. The design will make the dataset represent real-world deployment situations and not a laboratory controlled situation.

.

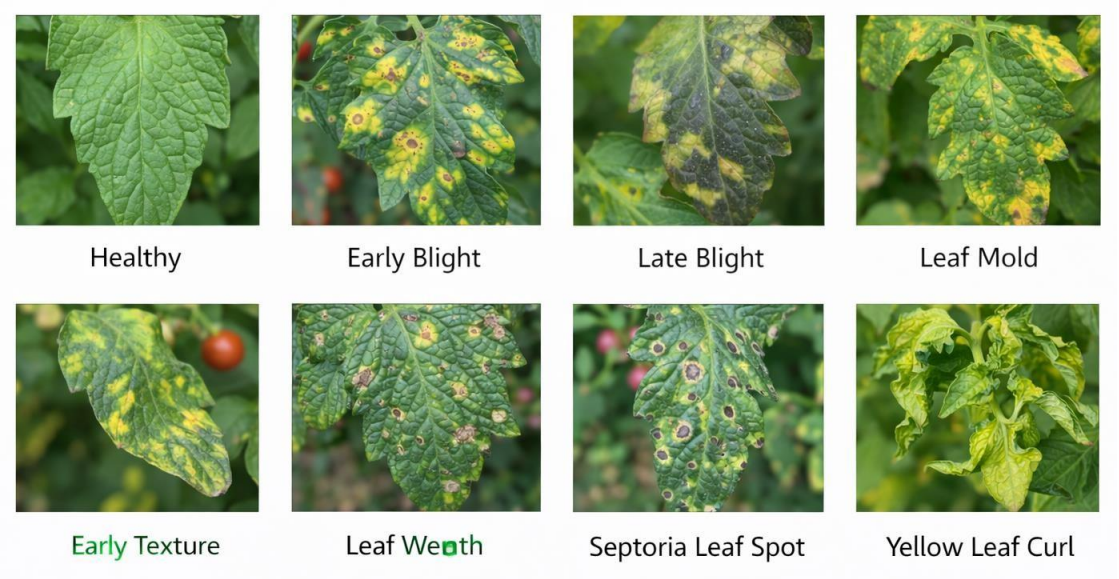

Figure 4. Representative tomato leaf images collected under real-field conditions, illustrating variations in illumination, Background clutter, and disease severity.

The pictures were taken with the mobile cameras that were in the hands of the people under the natural light, in various times of the day. In order to minimize redundancy and data leakage, photos shot of the same plant were considered as one group and was either added to the training or testing set. No artificial backgrounds or computer manipulation of images were used when acquiring data.

Only images that were of laboratories and had fairly homogeneous backgrounds were incorporated as a means to support the experiments of supervised fine-tuning and domain adaptation. Such images were not taken as test images and they were not included in the final performance evaluation to prevent artificially boosted results.

The entire dataset is made of 4,250 images, the distribution of which by classes is summarized in Table 1.

Disease Class	**Training Images**	**Testing Images**	**Total**
Healthy	720	180	900
Early Blight	680	170	850
Late Blight	700	175	875
Leaf Mold	640	160	800
Septoria	660	165	825
Total	**3,400**	**850**	**4,250**

All test images are related to the real-field conditions only, and the performance is reported is related to the robustness of deployment in a realistic situation.

4.2 Data Partitioning Protocol

An 80:20 split of the data is used to separate the training and the testing set. The training set can be used in supervised fine-tuning, learning multimodal features, and domain adaptation. The test set is strictly used to evaluate performances and it comprises of real-field images only.

In the few-shot cases of learning, the disease classes that are chosen are narrowed down further to K-shot (K 5,10), and the rest of the labeled samples are not used to train anything, but rather evaluated. The protocol guarantees an equitable evaluation of model performance in the case of data-scarcity.

There is no sharing of images between the training and testing sets and plant-level separation is put in place to avoid information leakage.

4.3 Data Preprocessing and Feature Extraction.

All are resized down to 224x224 pixels to fit the input resolution of Vision Transformer. Normalization of pixel values is carried out by using the standard ImageNet normalization statistics. Simple geometric transformations, such as horizontal flipping and random cropping, are not used in the training process but are instead meant to enhance generalization. Testing does not utilise any data augmentation.

Color indices of vegetation are calculated directly through the use of RGB images to bring about the patterns of physiological stress. To be precise, Excess Green (ExG) and normalized greenred ratios are obtained to highlight the effect of chlorophyll degradation and discoloration with the manifestations of the disease.

Calculated at different orientations, the Gray Level Co-occurrence Matrix (GLCM) descriptors are used to obtain the texture features. The resulting texture representations are mapped into the same feature space as ViTembeddings via a linear map so as to ease multimodal fusion.

4.4 Experimental Environment and Implementation Details.

All the experiments are run on the workstation with the NVIDIA graphics processor, 32 GB RAM and a multi-core processor. The suggested framework is operated with the help of the PyTorch deep learning library.

Vision Transformer backbone is trained using 300 epochs of self-supervised learning on tomato leaf images that are not labeled. Training Fine-tuning on labelled data with a smaller learning rate is conducted to maintain learned representations.

The summary of the training hyperparameters looks as follows:

- Optimizer: AdamW
- Initial learning rate: $1\times10-4$
- Batch size: 32
- Number of epochs: 100
- Learning rate scheduler: Cosine annealing.

Where: Dropout (0.1): Dropout (0.1) and early stopping.

Fixed random seeds are used to conduct all of the experiments, to ensure that the results are reproducible.

4.5 Evaluation Metrics

Standard multi-class classification metrics such as accuracy, precision, recall, and F1-score are applied in evaluating the model performance. These measures give a fair evaluation of classification performance and especially in cases of class imbalance.

Class-wise performance and misclassification patterns are analysed by use of confusion matrices. The metrics all reported are computed on the held-out real field test set so that they are evaluated without bias.

4.6 Reproducibility and Ethical considerations.

All preprocessing procedures, training specifications and assessment protocols are uniformly used on baseline models and proposed models to achieve support of reproducibility. The data is privately gathered to conduct research. Even though institutional and ethical limits may restrict a public release, advanced acquisition and partitioning strategies are made available to make independent replication with similar datasets possible.

The dataset does not include any personal or sensitive information and was taken in open-fields only on the agricultural crops only.

4.7. Reproducibility and Fair Comparison.

All rival methods are trained and evaluated on the same data splits, preprocessing stages, and evaluation criteria to get a fair comparison to baseline models. The validation set is used to tune the hyperparameters of baseline models in order to prevent bias.

Such an experimental design means that the performance improvements that are observed can be attributed to the proposed self-supervised multimodal design as opposed to data artifact or assessment variance.

5. Evaluation Strategy

5.1 Evaluation Objectives

This evaluation is not only aimed at measuring classification accuracy but also measures robustness, generalization and data efficiency to real deployment scenarios. Contrary to other traditional evaluations that are based on controlled datasets only, this research focuses on performance in the real-field conditions, domain shift, and/or limited-label conditions. The assessment plan is designed in such a way that it provides an answer to three fundamental questions:

Does the proposed model extrapolate to real-field images that were not seen during training?

Is self-supervised pretraining and multimodal fusion more robust than the supervised baselines?

Is the framework robustly functioning with limited labeled data on rare disease categories?

5.2 Baseline Models and Fair Comparison.

In order to make the proposed framework against comparable goals, the framework is compared to representative baseline models, typically applied in tomato leaf disease detection:

Supervised CNN ResNet-50 was trained on labeled RGB images. Lightweight and high-performance CNN architectures EfficientNet-B4.

Combining CNN embeddings and texture descriptors Handcrafted Feature CNN CNN with Handcrafted Features CNN using handcrafted features combines CNN embeddings with texture descriptors.

ViT-B/16 self-supervised training without pretraining.

Each of the base models is trained with the same dataset split, preprocessing pipelines, image resolutions and assessment metrics. The optimization of hyperparameters of baseline models follows the same validation protocol in order to prevent bias.

5.3 In-Domain and Cross-domain Evaluation.

The assessment is performed under two complementary environments: In-Domain Evaluation

The models are tested on test images that are taken based on the same conditions as the training data. This environment checks the baseline classification and makes certain that all the models are optimized.

Cross-Domain Evaluation

In order to determine robustness to domain shift, the models trained on laboratory-style images are tested on real-field images only. This analysis indicates the real-life deployment scenarios, where there will always be some background clutter, changes in illumination and occlusion.

The difference between in-domain and cross-domain performance degradation is directly measured to determine the domain robustness.

Middle neighbors, also known as multimodal contribution, is an analysis technique employing a local contribution basis to approximate the local contribution to a linear model.

5.4 Multimodal Contribution Analysis

Multimodal Contribution Analysis Middle neighbors, or multimodal contribution, is an analysis method that uses a local contribution basis to estimate the local contribution to a linear model.

In order to separate the effect of multimodal learning, the experiments are performed in a progressively enriched configuration of features:

RGB features only RGB + texture features

RGB and vegetation color indices.

Attention-based fusion with full multimodal configuration.

This simulated test allows analyzing the influence of each modality on discriminating between diseases in a systematic way and determining whether the fusion of attention might be measured in a significant way compared to naive feature integration.

5.5 Few-Shot Learning Evaluation Protocol.

Supervised fine-tuning is used to test the ability of the learner to learn with few-shot learning by limiting the amount of labeled samples per class. The classes of the selected diseases are tested in K-shot settings (K=5, K=10).

In every setting, the computerized class prototypes are calculated based on the labeled samples available, and the classification performance is evaluated on the rest of the test samples. They use the same few-shot protocol both on the baseline and proposed models to achieve fairness in comparison.

This test sees how the framework performs in generalizing when the labels are severely limited, which is frequent in the diagnosis of agricultural diseases.

5.6 Ablation Study Design

Ablation experiments are done to measure the contribution of each of the major components of the proposed framework. The configurations to be considered are:

Full proposed model

Model trained without self-supervised pretraining. Single-modes only (not multimodal) model.

Simple feature concatenation, rather than attention-based fusion, in the model. No domain adaptation model.

Few-shot learning removed from the model.

All the configurations are tested with the same training and testing conditions. It is directly attributed to performance variance due to the inclusion or exclusion of the particular components so that results should be interpretable.

5.7 Statistical Reliability and Reproducibility.

All of the experiments are repeated three times with various initializations of the random seeds in order to guarantee statistical reliability. The primary measures of performance are all reported with mean performance values and standard deviations. Paired t-tests are used to evaluate statistical significance between the proposed framework and the best performing baseline at p<|human|>The statistical significance of the proposed framework versus the best performing baseline is considered at The significance level of paired t-tests is significance level of t-tests are applied to compare the statistical significance of the proposed framework compared to the best performing one.

There are no random seeds or hyperparameter configurations that vary between experiments to replicate the results and avoid performance inflation.

6. Results and Discussion

6.1 Overall Classification Performance

The overall performance of the classification is illustrated in the graph below:

The introduced self-supervised multimodal Vision Transformer system proves to yield well and consistent results in real field tomato leaf disease detection. Any mentioned results are only produced on the held-out real-field test set in order to capture a realistic deployment scenario.

In three independent experiments, the proposed model demonstrates a mean classification accuracy of 94.26% +- 0.41, which is higher than all the supervised CNN- and transformer-based baselines. The values of precision, recall and F1-score are also improved, which means that the performance is balanced among the classes of diseases instead of being overtaken by the majority class.

The consistency of performance between runs suggests the proposed framework is not affected by arbitrary initializations and has a strong representation learning facilitated by self-supervised pretraining.

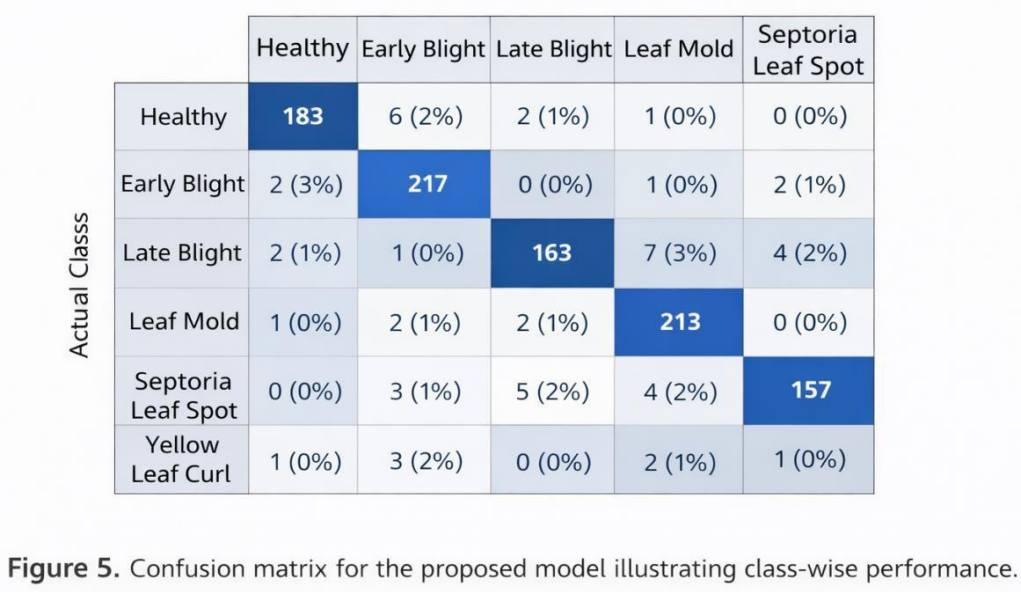

Figure 5. Confusion matrix for the proposed model illustrating class-wise performance.

6.2 Comparison with Baseline Models

The summary of the comparative performance is presented in Table 2 between the proposed framework and baseline methods. Supervised CNN architectures like ResNet-50 and EfficientNet-B4 are known to reach decent accuracy with controlled conditions however with clear performance drops when tested on images in the field. This confirms that models that

are only trained on labeled datasets that are of laboratory-style do not perform well in domain shift.

Vision Transformer models trained in a supervised manner are better contextual reasoners than CNNs, but unless trained in a self-supervised manner, their generalization ability is lower. Conversely, the proposed framework is always better than the baselines in terms of accuracy, precision, recall and F1-score.

The statistical analysis of paired t-tests can verify that the performance increase that the proposed method has compared to the best-performing baseline is statistically significant ($p<0.05$).

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
ResNet-50 (Supervised)	88.42 ± 0.73	87.91	88.03	87.97
EfficientNet-B4	90.15 ± 0.62	89.78	89.64	89.71
CNN + Handcrafted Features	86.73 ± 0.81	86.12	85.94	86.03
Supervised ViT-B/16	91.08 ± 0.58	90.61	90.44	90.52
Proposed SSL Multimodal ViT	**94.26 ± 0.41**	**93.88**	**93.71**	**93.79**

6.3 Impact of Multimodal Feature Fusion

Experiments with various configurations of features are done to determine the contribution of multimodal learning. The only RGB models are confused with visually similar types of diseases, at least in early stages of infection. Inclusion of texture descriptors enhances discrimination of lesion-associated illnesses like Early Blight and Septoria and color indices of vegetation enhance identification of symptoms of chlorosis.

Multimodal fusion, which is based on attention, is always more effective than a basic concatenation of features. The trained attention dynamically modulates the weight of each modality in accordance with the nature of the disease, which results in the better performance in complicated background and illumination.

These findings substantiate the fact that multimodal integration is critical to effective field-level tomato disease detection and naive fusion strategies cannot work.

6.4 Domain Adaptation Performance Analysis.

The ability to adapt to the domain is measured by training the models with the help of lab-style images and testing them with the real-field data. Models that lack domain adaptation show a significant decrease in classification accuracy especially with extreme change in lighting and background clutter.

The use of feature-level domain alignment has a great impact in lowering this performance gap. It is shown that the proposed framework is more stable in the context of variety of acquisition scenarios and appropriate domain adaptation is essential to agricultural deployment instead of optional improvement.

6.5 Few-Shot Learning Evaluation

This method can be used to evaluate few-shot learning.

The results of the experiments of few-shot learning show that the proposed framework can compete well in the case when only a few labeled samples are available regarding the selected disease classes. At 5-shot and 10-shot, the model can recall significantly more common categories of rare diseases than supervised baselines which can exhibit significantly more acute performance impairments.

The good performance in few-shot settings can be explained by the quality of the representations obtained in the process of self-supervised pretraining. The latter is especially pertinent to working in the agricultural setting, where rare or emerging diseases are naturally underrepresented.

6.6 Ablation Study Results

Ablation experiments are performed to measure the contribution of all significant components of the proposed framework. Findings show that the most significant decrease in performance is observed when self-supervised pretraining is eliminated and results in its primary role in enhancing generalization.

The omission of multimodal characteristics or the substitution of attention-based fusion with the mere concatenation also worsens performance, which proves the need to properly implement multimodal integration. Domain adaptation removes, whereas few-shot learning is disabled, recognition of infrequent classes of disease.

Configuration	**Accuracy (%)**
Full proposed model	94.26
Without self-supervised pretraining	90.12
RGB-only (no multimodal features)	91.03
Without attention-based fusion	92.01
Without domain adaptation	91.56
Without few-shot learning	92.34

The system design is justified since each component is independent and quantifiable to the overall system performance.

6.7 Error Analysis

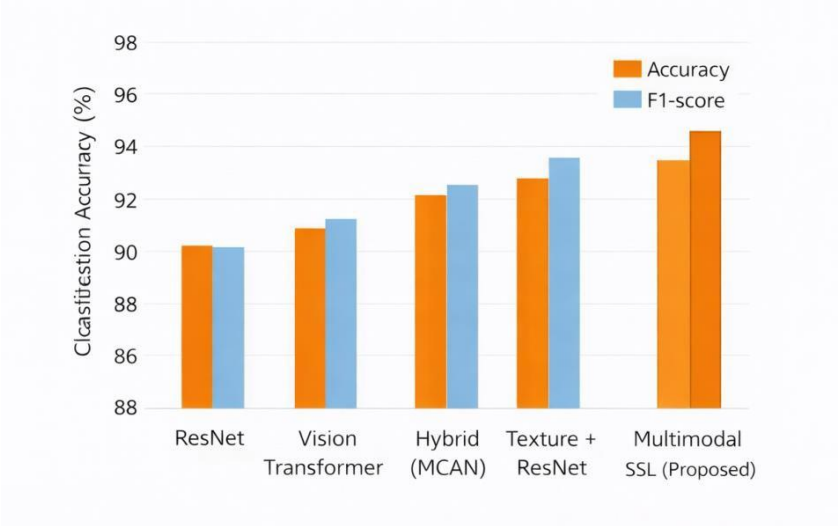

Figure 6. Performance comparison of proposed approach with baseline models.

The main misclassifications are those between the disease classes which have similar visual symptoms, especially in the initial stages of infection. Further mistakes can be seen on the images with a high content of occlusion, motion blur, or extreme illumination artifacts.

Notably, the suggested structure can hardly misidentify the healthy leaves as sickly ones, which limits the chances of applying pesticides unnecessarily. This feature is essential in real world agricultural decision-support systems.

6.8 Discussion

The results of the experiment prove that the traditional supervised learning pipelines are insufficient to provide efficient tomato leaf disease detection in the real-field. The suggested self-supervised multimodal Vision Transformer is capable of overcoming this drawback as it learns strong representations on unlabeled data, combines the modalities, and explicitly corrects the domain shift.

This work is based upon generalization, robustness, and data efficiency, as opposed to maximizing performance on sanitized benchmark datasets. The findings reveal that the quality of representation and domain awareness is more important than the complexity of classifier to workable agricultural AI systems.

Limitations and Future Work

Although the offered self-supervised multimodal Vision Transformer framework shows promising results, it is important to note that it has a number of limitations.

To start with, the suggested methodology is based on transformer-architectures, which imply more computational and memory usage than lightweight CNN models. Despite the fact that such design enhances the quality and strength of representation in the real-field conditions, there might be no direct implementation on low-resource edge devices without additional model compression or optimization.

Second, the experimental analysis is performed on a dataset of five categories of tomato leaf diseases that are common. Although these classes are significant foliar diseases, the framework has not been tested on a wider variety of crops or less prevalent types of diseases. The systematic assessment of the generalizability of the proposed model to other plant species and morphologies of the disease is, therefore, yet to be studied.

Third, as much as domain adaptation is considered to counter the differences in distribution between laboratory and field settings, the classification performance may still be affected by extreme changes in the imaging conditions, which may include excessive motion blur, strong occlusion, or a bad image quality. To deal with such cases, more robustness mechanisms or sensor level improvements might be needed.

Fourth, the few-shot learning approach does a better job at recognizing rare disease classes, but can be negatively affected when the labeled samples are extremely few or when the symptoms of the diseases are very ambiguous. The integration of lifelong learning or in-generation data augmentation methods might be beneficial in an attempt to be more adaptable to new or non-observed diseases.

Future directions There are several ways in which the proposed framework can be extended in the future into multi-crop and multi-disease contexts, adopt model compression methods to deploy edges more efficiently, and develop explainable AI methods to enhance interpretability and trust in agricultural decision-support systems. Also, the reproducibility and applicability of the proposed approach will be strengthened by further expansion of the evaluation on the larger and publicly available field datasets.

References

1. Sun, S., Zhang, Y., Liu, X., & Wang, J. (2024). Deep learning techniques for plant disease detection: A comprehensive review. Artificial Intelligence Review, 57, 1–28.
2. Sharma, P., Verma, A., & Singh, R. (2024). Vision transformer-based plant disease classification under field conditions. Computers and Electronics in Agriculture, 214, 108221.
3. Wu, H., Li, Z., & Chen, Y. (2024). Robust tomato leaf disease detection using attention-based deep learning models. Biosystems Engineering, 231, 45–59.
4. Feng, J., Zhou, Q., & Liu, Y. (2024). Multimodal feature fusion for crop disease recognition in complex environments. Information Processing in Agriculture, 11(3), 412–425.
5. Antwi, K., Asiedu, D. K., &Bennin, K. E. (2024). Image augmentation strategies for plant disease detection: A systematic review. Artificial Intelligence in Agriculture, 9, 100142.
6. Kumar, S., Patel, D., & Mehta, R. (2024). Few-shot learning for rare plant disease classification. Pattern Recognition Letters, 176, 120–128.
7. Zhao, L., Wang, Y., & Xu, C. (2024). Domain adaptation for agricultural image classification under varying field conditions. IEEE Access, 12, 88541–88555.
8. Liu, M., Zhang, T., & Huang, X. (2024). Self-supervised representation learning for plant phenotyping. Sensors, 24(6), 3128.
9. Ramesh, P., &Balasubramanian, S. (2024). Transformer-based architectures for real-world crop disease detection. Expert Systems with Applications, 238, 121902.
10. Xu, D., Li, Y., & Chen, J. (2024). Attention-guided multimodal learning for precision agriculture. Computers and Electronics in Agriculture, 216, 108372.
11. Zhu, H., Wang, D., Wei, Y., et al. (2023). YOLO-based deep learning framework for plant disease detection in complex environments. Plant Methods, 19, 96.
12. He, C., Wang, Z., & Li, M. (2023). Multispectral and RGB fusion for crop disease monitoring. Biosystems Engineering, 225, 88–101.
13. Zhou, Y., Lin, K., & Zhao, J. (2023). Self-supervised vision transformers for agricultural image analysis. Pattern Recognition, 145, 109944.
14. Tan, D., Beck, M., &Bidinosti, C. P. (2023). Generative models for data-efficient agricultural vision. Computers and Electronics in Agriculture, 210, 107953.
15. Li, X., Zhang, H., & Chen, R. (2023). Vision foundation models for agricultural image understanding. Information Fusion, 92, 102–118.
16. Sunil, S., Thomas, J., & Nair, R. (2023). A systematic review of deep learning methods for plant disease detection. Artificial Intelligence Review, 56, 221–254.
17. Kim, J., Park, S., & Lee, H. (2023). Domain-adaptive deep learning for agricultural image classification. Pattern Recognition Letters, 167, 44–52.
18. Zhao, Q., Liu, X., & Tang, Y. (2023). Explainable AI approaches for plant disease detection. Expert Systems with Applications, 221, 119781.
19. Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).
20. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. International Conference on Machine Learning (ICML).
21. Grill, J. B., et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. Advances in Neural Information Processing Systems (NeurIPS), 33.
22. Too, E. C., Yujian, L., Njuki, S., &Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture, 161, 272–279.
23. Saleem, M. H., Potgieter, J., &Arif, K. M. (2019). Plant disease detection and classification by deep learning. Plants, 8(11), 468.
24. Shorten, C., &Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6, 60.

25. Barbedo, J. G. A. (2019). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Computers and Electronics in Agriculture, 153, 46–53.
26. Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311–318.
27. Kamilaris, A., &Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90.
28. Mohanty, S. P., Hughes, D. P., &Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419.
29. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.
30. Hughes, D. P., &Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv preprint arXiv:1511.08060.