

Joint Plant Species-Disease Prediction: A Comparative Evaluation of Classification Approaches and Novel Conditional Multi-Task Learning Frameworks

Ameya Kella¹

Received April 19, 2026

Accepted July 13, 2026

Electronic access August 15, 2026

This study investigates automated plant species and disease classification using image-based machine learning models. Specifically, our analysis is based on the idea that disease prediction is inherently dependent on plant species. We propose two novel conditional multi-task learning (CMTL) frameworks, while comparing them to one-stage classification, two-stage classification, and prior fusion-based CMTL approaches. We utilize a diverse dataset of 40,000 images spanning 9 plant species, 20 unique disease categories, and 33 joint species-disease classes. All frameworks were trained across multiple random seeds under a unified protocol using fine-tuned transfer learning backbones and were evaluated using species-level, disease-level, and joint prediction metrics. The best one-stage and two-stage frameworks achieved joint Top-1 accuracies of 94.35% and 93.38%, respectively. Among the reimplementations of previously proposed CMTL frameworks, the highest joint Top-1 accuracy was 93.09%. Our proposed Conditional Gating (Mixture-of-Experts) and Species-Conditioned Attention frameworks both achieved a higher joint Top-1 accuracy of 97.98%, while also achieving higher true conjoint F1-scores commonly underperforming classes. These results suggest that explicitly modeling the dependency between plant species and diseases can improve joint prediction performance within the evaluated dataset setting. Despite high predictive performance, the proposed frameworks come with the trade-off of increased computational cost compared to other evaluated frameworks. Furthermore, the findings should be interpreted in light of several limitations. Evaluation was conducted on curated, publicly available data with moderate class imbalance mitigation. Future work should focus on external validation on additional datasets investigating causes behind certain underperforming classes.

Keywords: Plant, Species, Disease, Joint Species-Disease Prediction, Machine Learning (ML), Transfer Learning, One-Stage, Two-Stage, Multi-Task Learning (MTL), Conditional Multi-Task Learning (CMTL), Computer Vision

Introduction

Plant diseases pose a significant threat to food security and economic stability, causing an estimated annual loss in global food production of 10-16% and \$220 billion worth of economic damage¹. Detecting plant diseases earlier could minimize spread and thus prevent such large-scale losses. Traditional disease identification methods relied on manual disease diagnosis by experienced farmers or plant pathologists, but this is time-consuming and often unavailable in remote regions². Technological advancements have made machine learning and computer vision techniques feasible and practical alternatives for plant disease detection by shortening the diagnosis period³. Due to the widespread availability of digital cameras, smartphones, and drones, plant images can be readily captured at large scales and leveraged for automated disease classification using machine learning models⁴.

Many existing machine learning approaches can be grouped into one-stage, two-stage, and multi-task learning (MTL) frameworks. One-stage methods treat species-disease combinations as single classes, two-stage methods perform the two tasks sequentially, and MTL jointly predicts both outputs. However, MTL treats the outputs as parallel tasks, where species and disease predictions are separate. This motivates conditional multi-task learning (CMTL).

In this study, joint plant species-disease prediction is treated as a multi-output classification problem. Given an input image x , the objective is to predict both a species label (y_s) and a disease label (y_d), where y_s belongs to the set of plant species labels and y_d belongs to the set of disease labels. This dependency can be expressed as:

$$P(y_s, y_d | f(x)) = P(y_s | f(x))P(y_d | y_s, f(x))$$

where $f(x)$ denotes the shared feature representation extracted from the input image. This formulation reflects the

¹ Coppell High School, Texas, USA

hierarchical relationship between the tasks, where disease prediction is conditioned on species information. Accordingly, the learning objective is to jointly learn species and disease predictions while leveraging dependencies to improve overall prediction performance.

Although prior studies have compared one-stage, two-stage, and CMTL approaches, most have focused on a small set of fusion-based conditioning mechanisms. As a result, there is limited understanding of how alternative conditioning strategies perform under a unified dataset and training protocol. To address this gap, we propose two novel CMTL architectures that model species-disease dependencies using Conditional Gating and Species-Conditioned Attention. We then evaluate them against one-stage, two-stage, and existing fusion-based CMTL approaches.

Related Work

Research on automated crop disease detection has evolved significantly with advances in deep learning, transitioning from single-task classification models to more structured multi-task learning (MTL) frameworks. Table 1 summarizes representative studies that are most relevant to the methodological progression from single-task classification to conditional multi-task learning (CMTL). This overview highlights that prior work largely focuses on either single-task classification, shared multi-task prediction, or fixed hierarchical pipelines, without explicitly modeling conditional dependencies between plant species and disease classes. This gap motivates the approach proposed in this study.

Early work focused mainly on applying convolutional neural networks (CNNs) or transfer learning techniques on leaf images for plant disease classification. For example, an 8-layer CNN on tomato disease images and a pre-trained backbone on grapes both achieved high accuracy of 98.40%^{5,6}. However, these approaches are limited to single-task disease classification, requiring species-specific datasets and separately trained models that increase computational complexity. To address this, MTL frameworks have been introduced, in which a single model jointly predicts the species and disease. The PlantDisease Multi-task Joint Detection Model (PMJDM), for example, introduced a dual-branch architecture with shared feature extraction and task-specific heads, achieving competitive performance (61.83% mAP50) on multi-label agricultural datasets⁷. However, this approach relies on detailed multi-task annotations for each image, increasing dataset preparation costs. Other MTL-based studies have also explored attention mechanisms and Vision Transformer backbones for joint classification and localization tasks^{8,9}. While these methods improve feature sharing, they do not explicitly condition disease predictions on species. In parallel, hierarchical and structured prediction methods have been ex-

plored. One landmark work applied a one-stage classification approach on the PlantVillage dataset for 38 joint species-disease classes and achieved a high accuracy of 99.35% across multiple crops¹⁰. A two-stage classification approach decomposes the problem into sequential predictions, where species is predicted first, followed by disease classification, and this has achieved high results such as an average accuracy of 98.33%^{11,12}. While they reduce disease prediction complexity, these approaches require multiple models and separate training pipelines, which increases computational complexity by requiring multiple trained components. Importantly, they also do not enable end-to-end learning of species-disease interactions within a single unified model.

More general hierarchical learning techniques such as hierarchical softmax and label embedding have also been proposed in computer vision and large-scale classification tasks^{13,14}. These methods improve efficiency or capture semantic relationships between labels, but they primarily model static label structures rather than conditional dependencies between related prediction tasks. As a result, their direct application to plant species-disease prediction remains limited, particularly in tasks requiring explicit conditioning of disease prediction on species identity. Recent agricultural AI research has explored transformer-based architectures, self-supervised learning (SSL), and federated learning. Vision Transformers and attention-based models improve feature representation for plant disease recognition by capturing long-range dependencies in images¹⁵. Similarly, SSL approaches leverage large-scale unlabeled agricultural datasets to learn robust feature embeddings prior to fine-tuning¹⁶. Federated learning enables distributed training across decentralized agricultural data sources while preserving privacy¹⁷. Although beneficial for representation learning, data governance, and distributed model training, these methods are complementary to CMTL frameworks, which is a key focus of this study.

Finally, CMTL has emerged as a promising direction that explicitly models the dependency between the two outputs. Unlike standard MTL, CMTL introduces conditional mechanisms that allow disease prediction to be influenced by species information, thereby explicitly modeling dependencies between the two prediction tasks. Lee et al. compared one-stage, two-stage, and CMTL approaches using fusion-based conditioning mechanisms (FFS, FFC, and FiLM)¹⁸. Rather than introducing CMTL itself, the present study extends that line of work by proposing and evaluating alternative conditioning mechanisms (gating- and attention-based) under a different dataset configuration and unified training protocol.

Research Scope and Contributions

The scope of this study is limited to image-based plant species and disease classification using a subset of the PlantVillage¹⁰

Table 1 Representative studies in plant disease detection literature

Study	Model Type / Approach	Primary Dataset Images	Number of Species	Number of Diseases
Agarwal et al. (2020)	Lightweight, 8-layer CNN	18,000	1	9
Paymode and Malode (2022)	Transfer Learning VGG16 CNN (Implemented twice for 2 species)	18,483 (4,062 for Grape and 14,421 for Tomato)	2	12
Fu et al. (2025)	PMJDM (Multi-Task Learning Framework)	26,073 (with annotations for each)	5	28
Mohanty et al. (2016)	One-Stage Joint Classification	54,306	14	26
Costa et al. (2019)	Two-Stage Hierarchical CNN	23,988	3	13
Lee et al. (2021)	CMTL (FFS, FFC, and FiLM)	54,305	311	289

dataset containing 9 plant species and 20 disease categories. We limit class imbalance mitigation to only class-weighted loss functions, and we do not incorporate temporal, environmental, or genomic information. Although per-class results are provided, investigation of the causes of underperforming classes is left for future work. The study focuses on controlled dataset-based evaluation of MTL frameworks for joint plant species-disease prediction.

In this study, we propose two novel CMTL frameworks: Conditional Gating (Mixture-of-Experts) and Species-Conditioned Attention. We present a comparative evaluation of these frameworks alongside one-stage, two-stage, and reimplemented prior CMTL approaches using various pre-trained transfer learning backbones on a unified dataset.

Methods

In this study, multiple machine learning frameworks are trained and evaluated on the same dataset to enable comparison of model performance across different classification approaches.

Dataset

This study utilizes the PlantVillage dataset, a publicly available collection of labeled plant images for plant species and disease classification originally introduced by Mohanty et al.¹⁰. PlantVillage is a widely used benchmark dataset with expert-verified disease annotations. The dataset was downloaded on March 1, 2026 under the CC BY-SA 3.0 license from the official GitHub repository. Only the colored image subset was used in this work.

Because this study focuses on joint plant species-disease prediction, which inherently relies on intra-species class variability, species containing only a single disease class or only a “Healthy” class were excluded; this includes Orange, Blueberry, Raspberry, Soybean, and Squash. After this filtering, the final dataset consists of 40,000 images spanning 9 plant

species, 20 unique disease categories (including “Healthy”), and 33 joint species-disease classes. Figures 1 and 2 show image distributions across species and joint species-disease classes, respectively. Our dataset exhibits moderate class imbalance, with class sizes ranging from 152 images to 5,357 and a median class size of 1,052 images. Also, although acquisition metadata such as device type, geographic origin, and specific lighting conditions are not provided in the original dataset, all images were captured under varying conditions, resulting in differences in lighting and background. They originally varied in resolution, but Figure 3 shows representative sample images after resizing to 224×224.

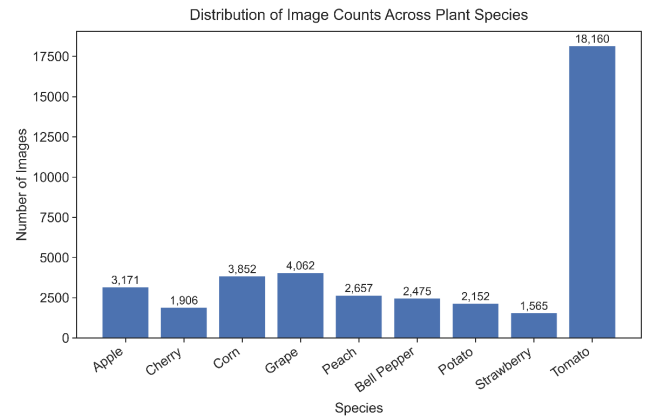


Figure. 1 Distribution of image counts across the 9 species in our dataset.

Data Preprocessing and Partitioning

We designed our preprocessing steps to ensure compatibility with our selected pre-trained backbones. For most experiments, images were resized to a consistent 224×224, while alternative resolutions were used when required by specific architectures. Pixel intensities were subsequently normalized

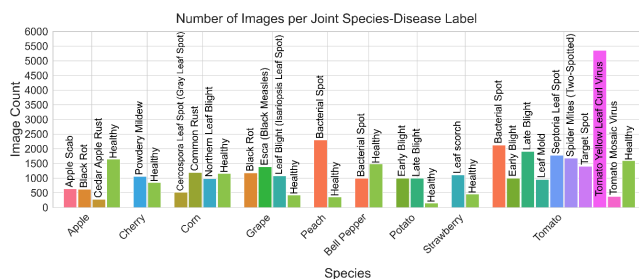


Figure. 2 Distribution of image counts across the 33 joint species-disease classes in our dataset.

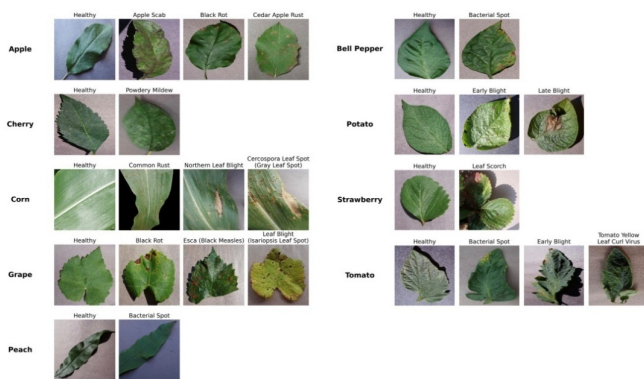


Figure. 3 Representative healthy and diseased images across multiple species-disease classes after resizing.

using backbone-specific preprocessing procedures. Furthermore, no data augmentation was applied during model training to isolate performance to model design rather than augmentation strategies. Also, our dataset contains 40,000 images spanning 33 species-disease classes, providing a sufficiently large dataset.

To reduce potential train-test leakage, a perceptual hash (pHash) analysis was performed. This identified 39,978 unique image groups, meaning there are 22 duplicate images (0.055% of the data). Dataset spitting was therefore performed at the duplicate-group level rather than at the individual-image level. Duplicate groups were stratified according to their joint species-disease labels and split into training (70%), validation (15%), and testing (15%) subsets using a random seed (42). This method ensures that visually identical images do not span multiple splits while maintaining joint class distributions for valid training and evaluation.

Descriptions of the Classification Approaches

In this study, we evaluate multiple approaches for joint plant species-disease prediction: one-stage, two-stage, and various

conditional multi-task learning (CMTL) frameworks.

One-Stage Classification Approach

First, we used the one-stage classification approach shown in Figure 4, in which one model was trained on all 33 joint species-disease classes. We leveraged VGG16, ResNet50, EfficientNetB0, and MobileNetV2 to fine-tune their learned features for species and disease classification^{19–22}. These models were selected due to their widespread use in plant disease detection literature^{23,24}.

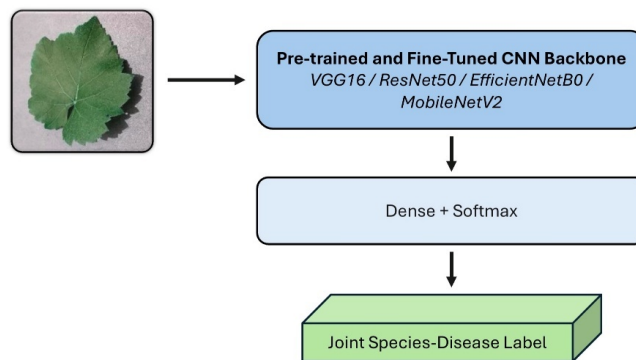


Figure. 4 Schematic representation of one-stage classification for joint species-disease prediction.

For each backbone, the original ImageNet²⁵ classification layer was replaced with a new head comprising global average pooling, dropout, and a 33-neuron softmax output layer.

Two-Stage Classification Approach

Next, as depicted in Figure 5, we used a two-stage classification approach in which plant species and disease classification are performed separately and sequentially. The same backbones and fine-tuning procedure used in the one-stage approach were applied to all models in this approach.

The first stage performs species classification using transfer learning. The second stage uses one of nine species-specific models, each trained only on diseases valid for that species. For inference, the first-stage species prediction determines which second-stage model is deployed.

Conditional Multi-Task Learning (CMTL) Approaches

To explicitly model relationships between plant species and diseases, we implemented several conditional multi-task learning (CMTL) approaches. Unlike the one-stage and two-stage approaches, CMTL jointly learns species and disease classification within a shared network while allowing species-related information to condition disease prediction.

Three frameworks were reimplemented from prior work

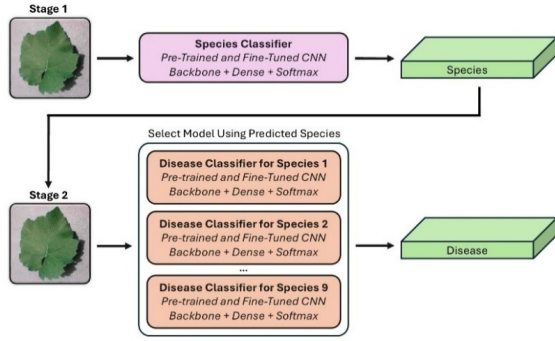


Figure. 5 Schematic representation of two-stage classification for joint species-disease prediction.

by Lee et al.¹⁸ and fine-tuned to our dataset: Feature Fusion Sum (FFS), Feature Fusion Cascade (FFC), and Feature-wise Linear Modulation (FiLM). These architectures used the pre-trained InceptionV3 backbone²⁶ to remain consistent with the original framework. In addition, two novel CMTL architectures were proposed in this study: a Conditional Gating Mixture-of-Experts framework and a Species-Conditioned Attention framework. Both proposed models employed a pre-trained VGG16 backbone¹⁹ due to its strong overall performance among the one-stage and two-stage approaches.

All CMTL models consisted of a shared feature extraction backbone followed by two task-specific branches: a species classification branch and a disease classification branch. The disease branch was conditioned using species-related information through various mechanisms.

Let h denote the backbone feature representation extracted from an input image x . The shared representation was projected into task-specific embeddings:

$$w_s(x) = \text{ReLU}(W_s h + b_s)$$

$$w_d(x) = \text{ReLU}(W_d h + b_d)$$

where $w_s(x)$ represents the species embedding and $w_d(x)$ represents the disease embedding, each containing 512 hidden units. Species classification was performed directly from $w_s(x)$, whereas disease classification depended on the framework-specific conditioning mechanism.

Reimplementation of Existing CMTL Architectures

Figure 6 shows the three CMTL fusion mechanisms proposed by Lee et al.¹⁸: Feature Fusion Sum (FFS), Feature Fusion Cascade (FFC), and Feature-wise Linear Modulation (FiLM).

In all cases, species and disease embeddings are combined to produce a species-conditioned representation that is subse-

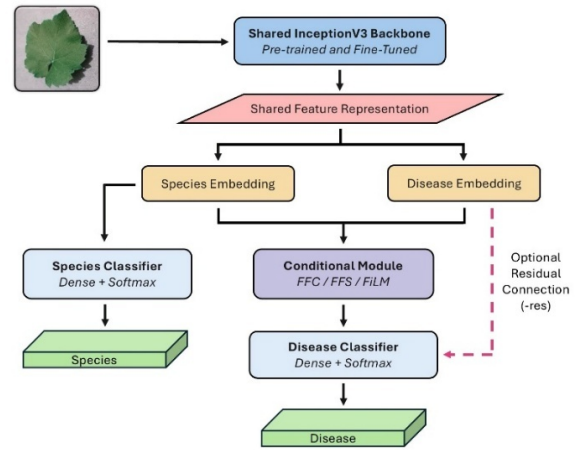


Figure. 6 Schematic representation of CMTL for joint species-disease prediction using FFS, FFC, and FiLM.

quently used for disease classification. All three architectures were reimplemented using the same progressive fine-tuning and hyperparameter selection strategy used throughout this study to ensure consistency. Based on the formulations proposed by Lee et al.¹⁸, the conditional fusion mechanisms are defined as follows.

The FFS architecture performs conditioning through additive feature fusion. Species and disease embeddings are first projected into a common latent space:

$$\sigma = H_\sigma(w_s(x))$$

$$\gamma = H_\gamma(w_d(x))$$

A conditioned representation is then computed using element-wise addition and linear activation:

$$F = \tanh(\sigma + \gamma)$$

The second reimplemented architecture utilizes cascade fusion (FFC), in which disease prediction is conditioned through feature concatenation rather than direct addition. The species and disease embeddings are first projected into a common feature space and concatenated:

$$F = \text{ReLU}(H_\beta([\sigma, \gamma]))$$

where $[\sigma, \gamma]$ denotes vector concatenation and H_β is a learned projection layer. Unlike FFS, the FFC model learns species-disease relationships through the concatenated representation.

The FiLM model employs a slightly different form of conditioning between the species and disease tasks, in which

species-conditioned scaling (σ) and shifting (β) parameters are learned from the species embedding:

$$\sigma = H_{\sigma}(w_s(x))$$

$$\beta = H_{\beta}(w_s(x))$$

and applied to a disease representation, which is computed as:

$$\gamma = H_{\gamma}(w_d(x))$$

The resulting conditioned representation is then computed as

$$F = \sigma \odot \gamma + \beta$$

where $\odot$ denotes element-wise multiplication¹⁸. Thus, disease features can be selectively amplified or suppressed based on species information^{18,27}.

For all three proposed architectures, Lee et al.¹⁸ also explored the addition of a residual connection (-res) in their analysis. Residual connections often improve gradient flow and stabilize training in deep neural networks²⁰. Because the three highest-performing architectures were FFS-res, FiLM-no-res (without a residual connection), and FFC-res, those same variants were reimplemented in this study. For the FFS and FFC models, the conditioned representation was combined with the original disease embedding through an identical mapping¹⁸:

$$y = \text{ReLU}(F + w_d(x))$$

where the residual connection reintroduces the original disease representation prior to disease classification. Consistent with the prior study's results, the FiLM framework in this study does not have a residual connection applied.

Proposed CMTL Framework 1: Conditional Gating (MoE)

Our first proposed CMTL framework is the Conditional Gating model incorporating a Mixture-of-Experts (MoE) architecture, as shown in Figure 7.

Rather than relying on a single disease classifier, we train multiple expert subnetworks to learn complementary disease representations. The motivation for this design originates from prior MoE research showing that expert specialization can improve modeling of heterogeneous prediction tasks^{28–30}. The gating network receives both the species embedding and species prediction probabilities. First, the gating logits (z) are computed as:

$$z = W_g[w_s(x), p_s(x)]$$

where $p_s(x)$ denotes the species prediction probabilities. A temperature-scaled softmax is then applied to obtain the expert-selection weights:

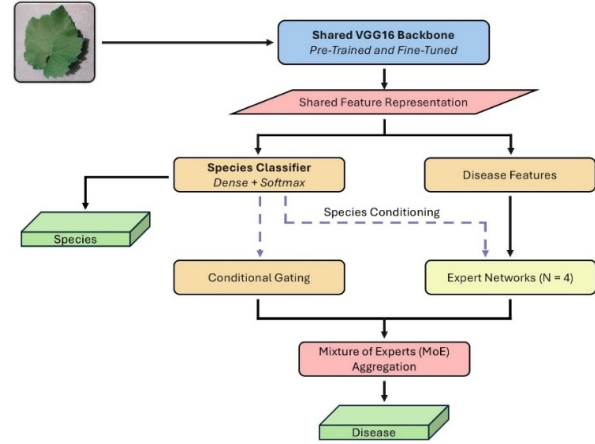


Figure. 7 Schematic representation of CMTL for joint species-disease prediction using Conditional Gating (MoE).

$$g = \text{softmax}\left(\frac{z}{T}\right)$$

where T denotes the temperature parameter. Temperature scaling moderates the sharpness of the expert-selection probabilities, with larger values producing smoother routing distributions^{29,31}.

The final disease representation is computed as a weighted combination of expert outputs:

$$F = \sum_{i=1}^N g_i E_i$$

where N denotes the number of experts. During development, multiple expert counts were evaluated: $N \in \{2, 4, 8, 16, 32\}$. We evaluated a logarithmic scale of powers of two to efficiently span a wide range of model capacities while minimizing hyperparameter optimization overhead. Based on validation disease loss, the final configuration of four experts was selected.

An entropy regularization term is also incorporated into the gating network and is minimized during optimization^{32–34}:

$$L_{\text{entropy}} = - \sum_{i=1}^N g_i \log(g_i)$$

which encourages lower-entropy gating distributions and more selective expert routing. This promotes stronger expert specialization during training by penalizing high-entropy gate outputs.

Proposed CMTL Framework 2: Species-Conditioned Attention (SCA)

Our second proposed CMTL architecture is the Species-Conditioned Attention (SCA) model shown in Figure 8. It

aims to allow species information to dynamically determine which disease features should receive greater emphasis during classification, motivated by the observation that certain visual disease symptoms may be more informative for specific plant species.

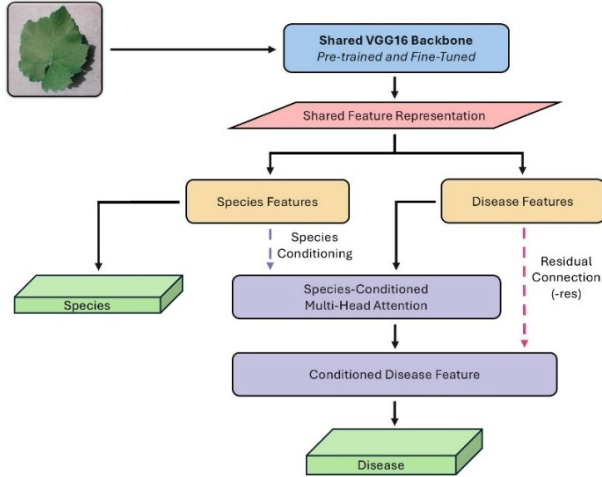


Figure. 8 Schematic representation of CMTL for joint species-disease prediction using SCA.

The species and disease embeddings were first projected into a shared latent space and represented as two feature tokens:

$$t_s = H_s(w_s(x))$$

$$t_d = H_d(w_d(x))$$

These tokens were concatenated into a two-token sequence,

$$T = [t_s, t_d]$$

which was then processed using multi-head attention (MHA) with four attention heads³⁵:

$$A = MHA(T, T, T)$$

A residual connection and layer normalization were subsequently applied to preserve disease information that may otherwise be weakened during attention-based modulation^{20,36}:

$$T' = \text{LayerNorm}(T + A)$$

The final conditioned disease representation was obtained from the updated disease token:

$$F = T'_d$$

where T'_d denotes the disease-token component of the normalized attention output.

Model Training and Fine-Tuning

Certain training settings were held constant across all frameworks. All models were initialized using backbones pre-trained on ImageNet²⁵ and subsequently fine-tuned on our dataset. Training was performed using the Adam optimizer with early stopping based on validation loss and ReduceLROnPlateau learning-rate scheduling. All experiments were trained for a maximum of 20 epochs on an NVIDIA GeForce RTX 4060 Laptop GPU. To quantify variability resulting from random initialization and stochastic optimization, every framework was trained using five random seeds (0, 1, 42, 123, and 492). Dropout was applied within the classification heads of the networks, and early stopping (with a patience of 3 epochs) served as an additional regularization mechanism. Weight decay was not employed.

Hyperparameters were selected through a framework-specific grid search. We evaluated learning rates of 10^{-3} , 10^{-4} , and 10^{-5} , batch sizes of 16 and 32, and dropout rates of 0.1, 0.2, and 0.3. To balance feature preservation and domain adaptation, we evaluated fine-tuning configurations with 2, 10, 20, or 40 trainable backbone layers depending on the architecture. Pre-trained backbones were also varied for the one-stage and two-stage approaches. Final model configurations were selected based on the lowest validation disease loss (for CMTL) or lowest overall validation loss (for one-stage and two-stage), maintaining consistency with the training optimization objectives.

Table 2 summarizes the final selected hyperparameter configurations and backbone choices for the best-performing version of each framework. These configurations were subsequently retrained across the five random seeds and were used for all reported results.

Table 2 Final selected hyperparameter configurations and backbone choices used for all reported results

Framework	Backbone	Learning Rate	Trainable Layers	Batch Size	Dropout
One-Stage	VGG16	10^{-5}	2	16	0.3
Two-Stage	VGG16	10^{-5}	2	16	0.3
CMTL – FFS	InceptionV3	10^{-5}	40	16	0.3
CMTL – FFC	InceptionV3	10^{-5}	40	16	0.3
CMTL – FiLM	InceptionV3	10^{-5}	40	16	0.3
CMTL – Gating	VGG16	10^{-5}	20	16	0.3
CMTL – SCA	VGG16	10^{-5}	20	16	0.3

For the Conditional Gating (MoE) architecture, additional framework-specific hyperparameters were explored. We evaluated expert counts of 2, 4, 8, 16, and 32, entropy regularization coefficients (or entropy weights) of 0.001, 0.005, and 0.01, and gating temperatures of 0.5, 1, and 1.5. Based on validation disease loss, the final configuration used four experts, an entropy regularization coefficient of 0.01, and a temperature of 1.5. A validation summary across expert-count config-

urations is provided as supplementary material.

Class imbalance was addressed through class-weighted loss functions. For one-stage classification, class weights were computed using the 33 joint species-disease classes. For two-stage classification, class weights were computed separately for the species classifier and for each species-specific disease classifier. For the conditional multi-task learning (CMTL) models, separate class-weight vectors were applied to the species and disease prediction tasks.

For all CMTL architectures, species prediction and disease prediction were learned jointly using a multi-task objective:

$$L = \alpha L_{\text{species}} + \beta L_{\text{disease}}$$

where L_{species} and L_{disease} are sparse categorical cross-entropy losses for species and disease classification, and α and β are the loss weights controlling the tasks' relative importance. We set $\alpha = \beta = 1$ in order to be consistent with Lee et al.¹⁸ and treat both tasks as primary objectives.

For the proposed Conditional Gating (MoE) architecture, an additional entropy regularization term was incorporated:

$$L = \alpha L_{\text{species}} + \beta L_{\text{disease}} + \gamma L_{\text{entropy}}$$

where γ denotes the entropy regularization coefficient (also referred to as the entropy weight). Because the entropy term is added to the training objective and is minimized during optimization, it encourages lower-entropy gating distributions. This promotes more selective expert routing and increased expert specialization.

Model Evaluation

To enable direct comparison across approaches, all models were evaluated using the same species-level, disease-level, and joint prediction metrics. Per-class precision, recall, and F1-scores were computed for all species, diseases, and joint species-disease classes. Additionally, we have computed the per-joint-label F1-scores, disease-only F1-scores conditional on correct species, disease-only F1-scores regardless of species, and true conjoint F1-scores. This is provided as supplementary material.

For species- and disease-level classification, precision, recall, weighted F1-score, macro F1-score, balanced accuracy, and Top-1 accuracy were computed. Weighted F1-score accounts for class frequency and provides an overall estimate of performance across the dataset. Macro F1-score assigns equal importance to each class and is therefore more sensitive to performance on classes with fewer samples. Balanced accuracy was additionally included to account for unequal class distributions. Because weighted averages can obscure poor performance on these underrepresented classes, both weighted and macro F1-scores are reported for all approaches.

For joint plant species-disease prediction, performance was evaluated using joint Top-1 accuracy. Joint Top-1 accuracy (or conjoint Top-1 accuracy) requires a model to correctly predict both the plant species and the disease label simultaneously for a given sample; an individual prediction is counted as correct only if both target labels match the ground truth. Additionally, joint-label performance was evaluated using weighted and macro F1-scores computed across the 33 joint species-disease classes. This ensured that all approaches were evaluated using a common prediction space.

To quantify uncertainty arising from random initialization, metrics are reported as mean $\pm$ 95% confidence interval across five independent training runs. Detailed per-seed and per-class results, as well as statistical testing outputs, are provided as supplementary material. Statistical significance between the five CMTL frameworks was assessed using pairwise McNemar tests on conjoint Top-1 predictions³⁷. For each model pair, accuracy differences ($\Delta = A - B$) were averaged across the five random seeds, and p-values were combined. Holm-Bonferroni correction³⁸ was applied as control for multiple comparisons within each task.

To compare computational costs, we recorded total trainable parameters, average training time (across the seeds), and average inference time per image. For the two-stage framework, parameters and training times are summed across its ten models. Its inference time equals the species classifier's inference time plus the mean inference time of the nine disease classifiers, as only the predicted species' disease classifier model is used.

Ethical Considerations

This study does not involve human participants or personal data. All images are from a publicly available dataset and were leveraged in accordance with the licensing and usage guidelines.

Results

This section evaluates all classification approaches explored in this study. These include a one-stage approach, a two-stage approach, three reimplemented CMTL frameworks (Feature Fusion Sum, FFS; Feature Fusion Cascade, FFC; and Feature-wise Linear Modulation, FiLM), and two novel CMTL frameworks (Conditional Gating and Species-Conditioned Attention, SCA).

Overall Predictive Performance Across Frameworks

Table 3 summarizes the predictive performance of all evaluated frameworks. This allows clear comparison at the joint-, species-, and disease- levels.

Table 3 Mean predictive performance ($\pm 95\%$ confidence interval across five random seeds) for all frameworks

Framework	One-Stage	Two-Stage	FFS	FFC	FiLM	Gating	SCA
Joint Top-1 Accuracy	0.9435 $\pm$ 0.0046	0.9338 $\pm$ 0.0055	0.9309 $\pm$ 0.0075	0.9281 $\pm$ 0.0086	0.9244 $\pm$ 0.0100	0.9798 $\pm$ 0.0066	0.9798 $\pm$ 0.0074
Joint Balanced Accuracy	0.9423 $\pm$ 0.0046	0.9153 $\pm$ 0.0060	0.9111 $\pm$ 0.0118	0.9098 $\pm$ 0.0107	0.9040 $\pm$ 0.0128	0.9728 $\pm$ 0.0083	0.9740 $\pm$ 0.0099
Joint Macro F1	0.9408 $\pm$ 0.0052	0.6075 $\pm$ 0.0210	0.4880 $\pm$ 0.0317	0.4749 $\pm$ 0.0317	0.4868 $\pm$ 0.0092	0.6821 $\pm$ 0.1287	0.6623 $\pm$ 0.0607
Joint Weighted F1	0.9439 $\pm$ 0.0043	0.9407 $\pm$ 0.0047	0.9384 $\pm$ 0.0074	0.9358 $\pm$ 0.0077	0.9325 $\pm$ 0.0104	0.9819 $\pm$ 0.0053	0.9817 $\pm$ 0.0069
Species Macro F1	0.9853 $\pm$ 0.0016	0.9671 $\pm$ 0.0079	0.9782 $\pm$ 0.0093	0.9794 $\pm$ 0.0026	0.9755 $\pm$ 0.0069	0.9907 $\pm$ 0.0063	0.9899 $\pm$ 0.0074
Species Weighted F1	0.9884 $\pm$ 0.0014	0.9705 $\pm$ 0.0068	0.9814 $\pm$ 0.0087	0.9828 $\pm$ 0.0022	0.9792 $\pm$ 0.0067	0.9924 $\pm$ 0.0052	0.9913 $\pm$ 0.0067
Disease Macro F1	0.9334 $\pm$ 0.0044	0.9491 $\pm$ 0.0043	0.9384 $\pm$ 0.0051	0.9352 $\pm$ 0.0081	0.9333 $\pm$ 0.0054	0.9826 $\pm$ 0.0018	0.9840 $\pm$ 0.0020
Disease Weighted F1	0.9479 $\pm$ 0.0039	0.9620 $\pm$ 0.0022	0.9471 $\pm$ 0.0046	0.9430 $\pm$ 0.0086	0.9417 $\pm$ 0.0051	0.9856 $\pm$ 0.0024	0.9874 $\pm$ 0.0018
ECE	0.0083 $\pm$ 0.0030	0.0115 $\pm$ 0.0056	0.0136 $\pm$ 0.0137	0.0110 $\pm$ 0.0054	0.0110 $\pm$ 0.0031	0.0051 $\pm$ 0.0024	0.0055 $\pm$ 0.0032

The one-stage framework outperformed the two-stage framework in joint Top-1 and balanced accuracy, but the two-stage approach showed better disease-level performance, as seen in the disease weighted F1-score. Among the three reimplementations of prior CMTL architectures, FFS achieved the highest conjoint Top-1 accuracy, followed by FFC. Species-level and disease-level metrics were generally similar across the three fusion-based architectures, although FFS consistently achieved the strongest overall performance within this group. The proposed Conditional Gating and SCA architectures achieved the strongest overall results across nearly all evaluation metrics. Conditional Gating achieved a conjoint Top-1 accuracy of $97.98\% \pm 0.66\%$, while SCA achieved $97.98\% \pm 0.74\%$. These values exceeded those of the one-stage, two-stage, and fusion-based CMTL approaches. Similar trends were observed for joint balanced accuracy and joint weighted F1-scores. Additionally, the gap between the joint macro F1-scores and joint weighted F1-scores across all frameworks suggests the presence of class imbalance despite our measures taken to mitigate it.

At the species level, both proposed architectures achieved weighted F1-scores above 0.99, exceeding all other frameworks. Disease-level performance followed a similar trend; SCA achieved the highest disease weighted F1-score (0.9874 ± 0.0018) and disease macro F1-score (0.9840 ± 0.0020), while Conditional Gating achieved nearly identical results. Regarding calibration performance, the Expected Calibration Error (ECE), was also strong across all frameworks, but the Conditional Gating and SCA frameworks achieved the lowest ECE values (0.0051 ± 0.0024 and 0.0055 ± 0.0032 , respectively). This indicates that their predicted probabilities were generally well aligned with observed outcomes in comparison to the other frameworks. Overall, the proposed CMTL architectures achieved the strongest predictive performance and the best calibration characteristics.

Computational Cost Analysis

Table 4 compares the computational characteristics of all frameworks for analyzing complexity.

The one-stage framework required the fewest trainable pa-

rameters and the shortest average training time, while the two-stage framework reflected much higher computational cost due to its reliance on 10 separately trained models. Among the reimplemented CMTL frameworks, the trainable parameter counts ranged from about 8.7 to 9.2 million. FFC and FiLM had the lowest inference times among all evaluated frameworks, requiring about 23 milliseconds per image, while FFS required 48.28. The proposed Conditional Gating and SCA frameworks required substantially more trainable parameters and longer training times than the other CMTL approaches. This introduces a trade-off for the proposed frameworks: despite high predictive performance across nearly all metrics, they also come with increased computational cost.

Statistical Significance Test

Figure 9 presents pairwise comparisons among the five CMTL frameworks using McNemar's test³⁷. Asterisks indicate statistical significance after correction ($\alpha=0.05$).

Among the reimplemented architectures, FFS generally achieved slightly higher performance than FFC and FiLM, although the magnitude of these differences was substantially smaller than the improvements associated with the proposed architectures. Across joint species-disease prediction, both proposed frameworks demonstrated accuracy improvements of approximately 0.05 over FFS, FFC, and FiLM. These differences were consistently identified as statistically significant across all five seed runs. In contrast, the comparison between Conditional Gating and SCA revealed only negligible performance differences. Although some pairwise comparisons reached statistical significance, the observed accuracy differences were extremely small relative to the differences observed between the proposed and reimplemented frameworks.

Commonly Underperforming Classes

To better understand performance at the disease and joint levels, we have additionally computed the per-joint-label F1-scores, disease-only F1-scores conditional on correct species, disease-only F1-scores regardless of species, and true conjoint F1-scores. This is provided as supplementary material. Here,

Table 4 Comparison of total computational cost across frameworks

Framework	One-Stage	Two-Stage	FFS	FFC	FiLM	Gating	SCA
Backbone	VGG16	VGG16	InceptionV3	InceptionV3	InceptionV3	VGG16	VGG16
Total Trainable Parameters	2,376,737	23,619,626	8,712,093	9,236,893	8,974,749	17,376,645	16,831,837
Per-Image Inference Time (ms)	29.03	33.24	48.28	23.1	23.06	8.13	46.91
Total Training Time (hrs)	0.7	1.44	0.87	0.79	0.77	3.92	4.24

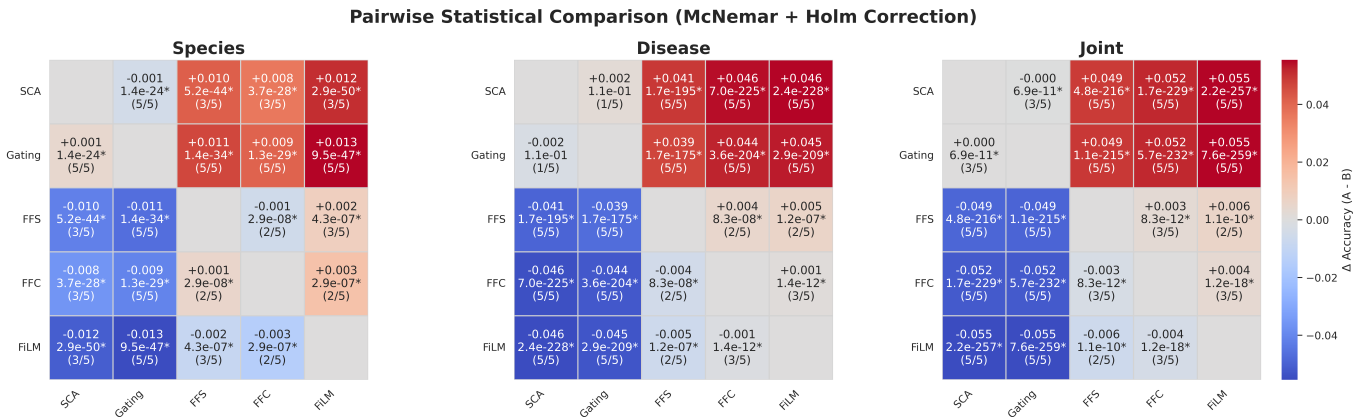


Figure. 9 Pairwise statistical comparison of CMTL frameworks using McNemar’s test.

Figure 10 presents true conjoint F1-scores for the lowest-performing joint species-disease classes after averaging across the five random seeds.

Several Tomato disease categories consistently appeared as the most challenging classes for all evaluated frameworks. This could be attributed to the large number of disease classes within Tomato species or subtle visual differences that are difficult for the models to distinguish between. The one-stage and two-stage frameworks generally exhibited the lowest performance on these classes, and the reimplemented CMTL frameworks often produced moderate improvements. On the other hand, the proposed Conditional Gating and SCA frameworks achieved the highest F1-scores across nearly all underperforming joint classes. For example, although these classes still performed worse than others, both approaches improved performance on Tomato’s Target Spot and Tomato Mosaic Virus relative to the other frameworks.

Discussion

This study evaluated various classification approaches for joint plant species-disease prediction under a unified experimental protocol, including two proposed conditional multi-task learning (CMTL) frameworks. The results provide insight into the advantages and trade-offs of explicitly modeling species-disease dependencies.

Key Findings and Interpretation

The results indicate that explicitly modeling species–disease dependencies can improve joint prediction performance relative to both one-stage and two-stage classification approaches. While the one-stage framework achieved strong overall performance and the two-stage framework produced competitive disease-level results, both approaches treat the relationship between species and disease either implicitly or through a fixed prediction pipeline. In contrast, the conditional multi-task learning frameworks incorporate species information directly into disease prediction, allowing the model to learn task interactions within a unified architecture.

Among the evaluated CMTL approaches, the proposed Conditional Gating and Species-Conditioned Attention models consistently achieved the strongest overall results, reaching conjoint Top-1 accuracies of approximately 97.98%, compared with 94.35% for the best one-stage model and 93.38% for the best two-stage model. They also achieved the highest disease-level F1-scores and the lowest calibration errors. These findings suggest that conditioning mechanisms based on dynamic expert selection or attention may capture species-specific disease characteristics more effectively than previously proposed fusion-based approaches.

Performance differences were particularly evident among the most challenging joint species–disease classes, many of which involved visually similar Tomato diseases. Although

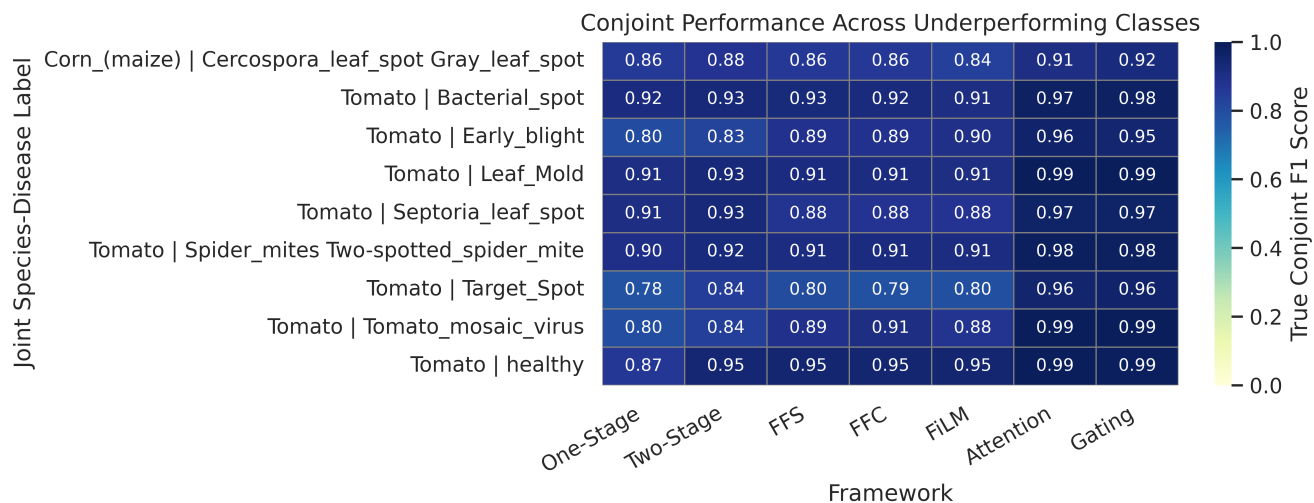


Figure. 10 Conjoint F1 performance across underperforming joint species-disease classes.

all frameworks exhibited worse performance on these classes, the proposed architectures generally showed greater robustness. This observation supports the hypothesis that explicit conditioning can help reduce ambiguity when disease symptoms share overlapping visual characteristics.

The computational analysis highlights an important trade-off. The proposed architectures achieved the highest predictive performance but required substantially more parameters and longer training times than simpler approaches. Therefore, the choice of framework may depend on whether predictive performance or computational efficiency is the priority.

Limitations and Future Work

Although this study is a comprehensive evaluation of several classification approaches for joint plant species-disease prediction, several limitations should be considered when interpreting the results. First, the models were evaluated only using stratified splits derived from a widely used, public dataset. Although the dataset includes 40,000 images captured under varying conditions, external validation using more test sets and a wider variety of species (and corresponding diseases) in future studies would allow for assessment on how well the models generalize to unseen acquisition settings, geographic regions, and field environments in the real world.

Second, the causes of poor performance for certain disease classes remain only partially understood. Future work may investigate feature-level confusion between visually similar diseases, quantify inter-class similarity, and analyze attention patterns or learned representations to better understand whether errors arise from dataset characteristics, label defini-

tions, or limitations of the model architectures themselves.

Additionally, moderate class imbalance remained present despite the use of class-weighted losses, as reflected by the gap between joint macro and weighted F1-scores. Future studies could explore alternative mitigation strategies. It may also consider incorporation of environmental, temporal, or genomic data to provide additional context for disease prediction. Finally, this study focused on comparing complete classification frameworks rather than isolating the contribution of individual conditioning components. Future work may include targeted ablation studies to further quantify the contribution of specific mechanisms.

Conclusions

This study compared one-stage classification, two-stage classification, and conditional multi-task learning (CMTL) approaches for joint plant species-disease prediction using a filtered subset of PlantVillage¹⁰ containing 40,000 images, 9 plant species, 20 disease categories, and 33 joint species-disease classes. All frameworks were evaluated under a common experimental protocol. The results show that explicitly modeling the conditional relationship between species and disease improves predictive performance. The proposed Conditional Gating and Species-Conditioned Attention frameworks achieved the strongest overall results, outperforming both conventional classification approaches and previously proposed fusion-based CMTL frameworks across joint, species-level, and disease-level evaluation metrics.

More broadly, the findings suggest that conditional learning can be beneficial in multi-task problems where one pre-

diction task naturally constrains another. However, the results were obtained on a curated benchmark dataset without external validation and should therefore be interpreted within that context. Future work should evaluate these architectures on larger and more diverse field-collected datasets and further investigate the sources of error among visually similar disease classes.

Overall, this study demonstrates the potential of CMTL for joint plant species-disease prediction and provides preliminary evidence that gating- and attention-based conditioning mechanisms are promising alternatives to traditional fusion-based approaches.

Code Availability and Supplemental Reports

The accompanying supplementary repository provides training scripts for each framework, dataset metadata, split mappings, per-seed and per-class evaluation metrics, a validation summary across different Conditional Gating models (varying expert counts), and the full McNemar test outputs. This is located at <https://github.com/ameyakella/JointPlantSpeciesDiseasePredCMTL>.

Acknowledgements

Although this paper uses the “authorial we” common in academia, all of the research, analysis, and writing were done solely by the author. I would like to thank Abdulla Kerimov (PhD in Geophysics from Stanford, Applied Scientist at BP) for guiding and supporting me throughout the process.

References

- 1 S. Chakraborty and A. C. Newton, *Climate Change, Plant Diseases and Food Security: An Overview*. *Plant Pathology*. Vol. 60, No. 1, pp. 2–14, 2011., <https://doi.org/10.1111/j.1365-3059.2010.02411.x>.
- 2 B. Sambana and et al., *An Efficient Plant Disease Detection Using Transfer Learning Approach*. *Scientific Reports*. Vol. 15, No. 1, p. 19082, 2025., <https://doi.org/10.1038/s41598-025-02271-w>.
- 3 T. Nyawose and R. C. Maswanganyi and P. Khumalo, *A Review on the Detection of Plant Disease Using Machine Learning and Deep Learning Approaches*. *Journal of Imaging*. Vol. 11, No. 10, p. 326, 2025., <https://doi.org/10.3390/jimaging11100326>.
- 4 N. Raji, *Plant Disease Classification Using Mobile-Captured Images: A Deep Learning Approach*. *International Journal of Intelligent Systems and Applications in Engineering*. Vol. 12, No. 3, pp. 3575–3581, 2024.
- 5 M. Agarwal and S. Kr. Gupta and K. K. Biswas, *Development of Efficient CNN Model for Tomato Crop Disease Identification*. *Sustainable Computing: Informatics and Systems*. Vol. 28, p. 100407, 2020., <https://doi.org/10.1016/j.suscom.2020.100407>.
- 6 A. S. Paymode and V. B. Malode, *Transfer Learning for Multi-Crop Leaf Disease Image Classification Using Convolutional Neural Network VGG*. *Artificial Intelligence in Agriculture*. Vol. 6, pp. 23–33, 2022., <https://doi.org/10.1016/j.aiia.2021.12.002>.
- 7 R. Fu and X. Wang and S. Wang and H. Sun, *PMJDM: A Multi-Task Joint Detection Model for Plant Disease Identification*. *Frontiers in Plant Science*. Vol. 16, p. 1599671, 2025., <https://doi.org/10.3389/fpls.2025.1599671>.
- 8 D. Dai and P. Xia and Z. Zhu and H. Che, *MTDL-EPDCLD: A Multi-Task Deep-Learning-Based System for Enhanced Precision Detection and Diagnosis of Corn Leaf Diseases*. *Plants*. Vol. 12, No. 13, p. 2433, 2023., <https://doi.org/10.3390/plants12132433>.
- 9 S. Hemalatha and J. J. B. Jayachandran, *A Multitask Learning-Based Vision Transformer for Plant Disease Localization and Classification*. *International Journal of Computational Intelligence Systems*. Vol. 17, No. 1, p. 188, 2024., <https://doi.org/10.1007/s44196-024-00597-3>.
- 10 S. P. Mohanty and D. P. Hughes and M. Salathé, *Using Deep Learning for Image-Based Plant Disease Detection*. *Frontiers in Plant Science*. Vol. 7, p. 1419, 2016., <https://doi.org/10.3389/fpls.2016.01419>.
- 11 J. Costa and C. Silva and B. Ribeiro, *Hierarchical Deep Learning Approach for Plant Disease Detection*. In: *Pattern Recognition and Image Analysis (Lecture Notes in Computer Science)*. Vol. 11868, pp. 383–393, 2019., https://doi.org/10.1007/978-3-030-31321-0_33.
- 12 T. T. Showrav and S. Bain and M. Hossain and K. I. Ahmed and S. A. Fattah and S. Ahmed, *A Two-Stage Approach for Plant Disease Classification Based on Deep Neural Networks and Transfer Learning*. In: *2022 12th International Conference on Electrical and Computer Engineering (ICECE)*, pp. 469–472, 2022., <https://doi.org/10.1109/ICECE57408.2022.10088587>.
- 13 A. A. Mohammed and V. Umaashankar, *Effectiveness of Hierarchical Softmax in Large Scale Classification Tasks*. In: *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1090–1094, 2018., <https://doi.org/10.1109/ICACCI.2018.8554637>.
- 14 Z. Akata and F. Perronnin and Z. Harchaoui and C. Schmid, *Label-Embedding for Image Classification*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 38, No. 7, pp. 1425–1438, 2016., <https://doi.org/10.1109/TPAMI.2015.2487986>.
- 15 S. Barath and S. M., *Federated Transformer–Blockchain Framework for Secure and Generalized Crop Disease Detection in Smart Agriculture*. *Journal of the Saudi Society of Agricultural Sciences*. Vol. 25, No. 4, p. 59, 2026., <https://doi.org/10.1007/s44447-026-00153-9>.
- 16 A. A. Mamun and M. Zhang and D. Ahmedt-Aristizabal and Z. Hayder and M. Awrangjeb, *StateSpace-SSL: Linear-Time Self-Supervised Learning for Plant Disease Detection*. *arXiv:2512.09492*, 2025., <https://arxiv.org/abs/2512.09492>.
- 17 D. Mamba Kabala and A. Hafiane and L. Bobelin and R. Canals, *Image-Based Crop Disease Detection with Federated Learning*. *Scientific Reports*. Vol. 13, No. 1, p. 19220, 2023., <https://doi.org/10.1038/s41598-023-46218-5>.
- 18 S. H. Lee and H. Goeau and P. Bonnet and A. Joly, *Conditional Multi-Task Learning for Plant Disease Identification*. In: *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3320–3327, 2021., <https://doi.org/10.1109/ICPR48806.2021.9412643>.
- 19 K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*. *arXiv:1409.1556*, 2014., <https://arxiv.org/abs/1409.1556>.
- 20 K. He and X. Zhang and S. Ren and J. Sun, *Deep Residual Learning for Image Recognition*. *arXiv:1512.03385*, 2015., <https://arxiv.org/abs/1512.03385>.
- 21 M. Tan and Q. V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. *arXiv:1905.11946*, 2019., <https://arxiv.org/abs/1905.11946>.
- 22 M. Sandler and A. Howard and M. Zhu and A. Zhmoginov and L.-C. Chen, *MobileNetV2: Inverted Residuals and Linear Bottlenecks*.

-
- arXiv:1801.04381, 2018., <https://arxiv.org/abs/1801.04381>.
- 23 G. S. Hukkeri and B. C. Soundarya and H. L. Gururaj and V. Ravi, *Classification of Various Plant Leaf Disease Using Pretrained Convolutional Neural Network on ImageNet*. *Open Agriculture Journal*. Vol. 18, No. 1, p. e18743315305194, 2024., <https://doi.org/10.2174/0118743315305194240408034912>.
 - 24 T. S. Alam and C. B. Jowthi and A. Pathak, *Comparing Pre-trained Models for Efficient Leaf Disease Detection: A Study on Custom CNN*. *Journal of Electrical Systems and Information Technology*. Vol. 11, No. 1, p. 12, 2024., <https://doi.org/10.1186/s43067-024-00137-1>.
 - 25 J. Deng and W. Dong and R. Socher and L.-J. Li and K. Li and L. Fei-Fei, *ImageNet: A Large-Scale Hierarchical Image Database*. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009., <https://doi.org/10.1109/CVPR.2009.5206848>.
 - 26 C. Szegedy and V. Vanhoucke and S. Ioffe and J. Shlens and Z. Wojna, *Rethinking the Inception Architecture for Computer Vision*. arXiv:1512.00567, 2015., <https://arxiv.org/abs/1512.00567>.
 - 27 E. Perez and F. Strub and H. de Vries and V. Dumoulin and A. Courville, *FiLM: Visual Reasoning with a General Conditioning Layer*. arXiv:1709.07871, 2017., <https://arxiv.org/abs/1709.07871>.
 - 28 R. A. Jacobs and M. I. Jordan and S. J. Nowlan and G. E. Hinton, *Adaptive Mixtures of Local Experts*. *Neural Computation*. Vol. 3, No. 1, pp. 79–87, 1991., <https://doi.org/10.1162/neco.1991.3.1.79>.
 - 29 N. Shazeer and et al., *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*. arXiv:1701.06538, 2017., <https://arxiv.org/abs/1701.06538>.
 - 30 J. Ma and Z. Zhao and X. Yi and J. Chen and L. Hong and E. H. Chi, *Modeling Task Relationships in Multi-Task Learning with Multi-Gate Mixture-of-Experts*. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1930–1939, 2018., <https://doi.org/10.1145/3219819.3220007>.
 - 31 C. Guo and G. Pleiss and Y. Sun and K. Q. Weinberger, *On Calibration of Modern Neural Networks*. arXiv preprint arXiv:1706.04599, 2017., <https://arxiv.org/abs/1706.04599>.
 - 32 G. Pereyra and G. Tucker and J. Chorowski and Ł. Kaiser and G. Hinton, *Regularizing Neural Networks by Penalizing Confident Output Distributions*. arXiv preprint arXiv:1701.06548, 2017., <https://arxiv.org/abs/1701.06548>.
 - 33 B. Peralta and A. Saavedra and L. Caro and A. Soto, *Mixture of Experts with Entropic Regularization for Data Classification*. *Entropy*. Vol. 21, No. 2, p. 190, 2019., <https://doi.org/10.3390/e21020190>.
 - 34 X. Chen and J. Zhou and Y. Li and J. Chen and S. Yu and Q. Xuan, *Mixture of Message Passing Experts with Routing Entropy Regularization for Node Classification*. arXiv preprint arXiv:2502.08083, 2025., <https://arxiv.org/abs/2502.08083>.
 - 35 A. Vaswani and N. Shazeer and N. Parmar and J. Uszkoreit and L. Jones and A. N. Gomez and Ł. Kaiser and I. Polosukhin, *Attention Is All You Need*. arXiv preprint arXiv:1706.03762, 2017., <https://arxiv.org/abs/1706.03762>.
 - 36 J. L. Ba and J. R. Kiros and G. E. Hinton, *Layer Normalization*. arXiv preprint arXiv:1607.06450, 2016., <https://arxiv.org/abs/1607.06450>.
 - 37 Q. McNemar, *Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages*. *Psychometrika*. Vol. 12, No. 2, pp. 153–157, 1947., <https://doi.org/10.1007/BF02295996>.
 - 38 S. Holm, *A Simple Sequentially Rejective Multiple Test Procedure*. *Scandinavian Journal of Statistics*. Vol. 6, No. 2, pp. 65–70, 1979., <https://www.jstor.org/stable/4615733>.