



Two-stage Transfer Learning for Chilli Leaf Disease Classification using CNN and Vision Transformer

Rudi Kurniawan¹, Samsuryadi Samsuryadi², Fatma Susilawati Mohamad³, Bunga Intan⁴, Nelly Khairani Daulay⁵

^{1,4}Department of Informatics, Faculty of Engineering Science, Universitas Bina Insan, Lubuk Linggau, Indonesia

²Department of Informatics Engineering, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

³Department of Information Technology, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Kuala Terengganu, Malaysia

⁵Department of Computer System Engineering, Faculty of Engineering Science, Universitas Bina Insan, Lubuk Linggau, Indonesia

*rudi.kurniawan@univbinainsan.ac.id

Abstract

Chili cultivation is considerably affected by leaf diseases, making timely and reliable disease identification essential for maintaining agricultural productivity. While deep learning has demonstrated strong capabilities in plant disease recognition, many existing studies primarily focus on modifying network architectures rather than optimizing the training process. This study introduces a two-stage transfer learning framework that combines frozen-backbone adaptation with controlled fine-tuning to enhance classification performance. The experiments utilized a publicly available dataset comprising 12,000 augmented chili leaf images representing six disease categories. Several deep learning architectures were investigated, including DenseNet121, EfficientNetB0, MobileNetV2, ResNet152V2, and the Vision Transformer (ViT-B16). The experimental results indicate that the proposed two-stage training strategy improves classification performance consistently across the evaluated architectures. Among the models, ViT-B16 obtained the highest test accuracy of 99.61%, together with stable macro-averaged precision, recall, and F1-score values. Furthermore, the confusion matrix and ROC analysis revealed strong discrimination among the disease classes, with only limited errors occurring between visually similar categories. These findings highlight the importance of optimizing the training schedule when adapting pretrained models for plant disease recognition. Overall, the proposed framework provides an effective and scalable approach for chili leaf disease classification while offering an alternative to approaches that primarily rely on architectural modifications.

Keywords: convolutional neural network; plant disease classification; precision agriculture; transfer learning; vision transformer

How to Cite: R. Kurniawan, S. Samsuryadi, F. S. Mohamad, B. Intan, and N. K. Daulay, "Two-stage Transfer Learning for Chilli Leaf Disease Classification using CNN and Vision Transformer", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1328–1339, Oct. 2026, doi: [10.29207/resti.v10i5.7674](https://doi.org/10.29207/resti.v10i5.7674).

Permalink/DOI: <https://doi.org/10.29207/resti.v10i5.7674>

Received: March 29, 2026

Accepted: September 5, 2026

Available Online: October 4, 2026

This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Published by [Ikatan Ahli Informatika Indonesia](https://www.ikatanahliinformatika.org/)

1. Introduction

Horticultural agriculture, particularly the cultivation of chili pepper (*Capsicum annuum*), plays an important role in the global food system. The sustainability of chili production is highly dependent on plant health, where leaf condition serves as a primary indicator of disease. Delayed disease detection can lead to rapid infection spread, increased control costs, and significant reductions in yield quality and quantity. By automatically analyzing visual symptoms, deep learning models can facilitate more efficient disease diagnosis than conventional approaches based on manual image inspection [1]. However, the complexity

and variability of disease symptoms remain challenging for automated classification systems.

From a biological perspective, chili leaf diseases exhibit considerable variability in their causes, manifestations, and spatial distributions. Chili leaf curl disease, which is associated with begomoviruses, demonstrates substantial genetic diversity and wide geographical distribution [2]. In addition, emerging pathogens such as *Fusarium proliferatum* further contribute to the complexity of disease patterns [3]. These conditions may produce heterogeneous symptoms, including deformation, discoloration, and necrosis, which can occur across different regions of the leaf. Therefore,

reliable classification requires visual representations capable of capturing both discriminative local characteristics and broader spatial relationships.

Deep learning has increasingly been adopted for automated plant disease recognition, with CNN-based models playing an important role in extracting discriminative patterns from leaf images. A range of pretrained networks, including ResNet, DenseNet, EfficientNet, and MobileNet, has been investigated for chili leaf disease identification and has produced promising classification results [4], [5], [6]. The applicability of these models extends beyond experimental evaluation to mobile-based diagnostic systems. For example, a CNN-driven mobile application achieved 93.31% accuracy in distinguishing healthy leaves from unhealthy ones and 99.03% when identifying six disease categories [7]. In another comparative study involving VGG16, VGG19, InceptionV3, and AlexNet, InceptionV3 obtained a 93% accuracy rate across four chili disease classes [8]. Image-based classification has also been explored using traditional machine learning techniques; an SVM-based approach reached 95% accuracy after incorporating image segmentation and post-processing procedures [9]. Nevertheless, CNN models offer a distinct methodological advantage because their feature representations are learned automatically from the input images according to the classification objective, reducing the dependence on manually engineered features.

Subsequent studies have improved CNN performance through architectural modification, ensemble learning, and data augmentation. Incorporating squeeze-and-excitation mechanisms into dense CNN architectures resulted in 97% accuracy, increasing to 98.86% after augmentation [10]. Residual Nested Dilated DenseNet (RNDDNet) achieved 98.09% accuracy, with precision and F1-score values of 97% and 97.25%, respectively [11]. Ensemble learning has also produced competitive results, with Chilli-Net, which combines ResNet50, VGG16, and InceptionV3, achieving 97% accuracy and an AUC of 100% [12]. Modified MobileNet achieved 97.18% accuracy while emphasizing computational efficiency [13], whereas data augmentation improved accuracy from 86% to 95.89% [14]. Pretrained EfficientNet-B0 achieved 97.05% accuracy under five-fold cross-validation [15], while an ensemble of EfficientNetV2B0 and EfficientNetB7 reached 99.56% accuracy [16].

Research has also extended beyond laboratory-style classification toward deployment and field-oriented scenarios. YOLOv5 achieved a mean Average Precision (mAP) of 0.461 for chili disease detection in unstructured field environments [17]. RegNet optimization resulted in 97.65% accuracy with reduced parameter complexity [18], while a GoogLeNet model optimized using a hybrid EGME0 strategy achieved 92.8% accuracy [19]. On a field-acquired chili dataset, fine-tuned DenseNet121 achieved 98.7% accuracy,

compared with 96.2% for ResNet-50, 95.1% for VGG16, and 97.3% for EfficientNetB4 [20]. These results demonstrate the maturity of CNN-based approaches for chili leaf disease recognition. Nevertheless, high classification performance alone does not resolve the fundamental limitation of convolution-based representation.

CNNs construct visual representations primarily through localized convolutional operations. Although receptive fields expand hierarchically with network depth, the representation process remains fundamentally dependent on local feature aggregation [21]. Consequently, spatial relationships between distant regions of an image are not modeled explicitly. This characteristic may become problematic for chili leaf diseases whose symptoms can occur as irregularly distributed regions across the leaf surface. Increasing network depth or architectural complexity may provide broader contextual information but can also increase computational requirements and training complexity.

Vision Transformer (ViT) architectures provide an alternative representation mechanism based on global self-attention. In the ViT architecture, the input image is represented as multiple image patches, whose contextual dependencies are captured by a self-attention mechanism that enables information exchange across the entire patch set [22]. This mechanism enables the explicit modeling of long-range spatial dependencies and may therefore be advantageous for disease symptoms distributed across multiple regions of a leaf. Initial applications of ViT to chili disease classification have shown promising results; the ChilliLeaf-ViT model, for example, achieved 99.38% accuracy across six chili leaf disease classes [23].

Despite these promising results, an important research gap remains. Existing studies have predominantly evaluated CNN or ViT architectures independently, often using different datasets, preprocessing procedures, training configurations, and evaluation protocols. Consequently, direct conclusions regarding the relative behavior of convolution-based and self-attention-based representations remain difficult to establish. Systematic comparisons involving multiple representative CNN architectures and ViT models under the same experimental conditions remain limited. Moreover, reported improvements are frequently attributed to architectural modifications, while the influence of the transfer learning strategy receives comparatively less attention.

Transfer learning is particularly important in agricultural image classification because pretrained models are commonly initialized using ImageNet, whose images differ substantially from domain-specific leaf images in terms of texture, color distribution, illumination, background, and structural characteristics. Direct fine-tuning of all pretrained parameters may therefore not provide an optimal adaptation pathway for the target domain. A staged adaptation strategy can provide a more controlled transition from general-

purpose visual representations to domain-specific representations.

A gradual adaptation procedure based on transfer learning is introduced to address this limitation [24-26]. Rather than updating the entire pretrained model from the outset, the initial phase keeps its backbone parameters fixed and trains the classification head for the target chili leaf disease categories. Subsequently, selected pretrained layers are released for further optimization so that the extracted features can be adjusted to characteristics specific to the target dataset. Such a sequential training scheme promotes controlled adaptation and helps preserve the general visual knowledge encoded in the original pretrained network.

Based on these gaps, this study establishes a unified experimental framework to compare convolution-based and self-attention-based visual representations for chili leaf disease classification. Four representative CNN architectures, namely DenseNet, MobileNet, EfficientNet, and ResNet, are evaluated alongside Vision Transformer models using the same dataset, preprocessing procedure, data partitioning, training configuration, two-stage transfer learning strategy, and evaluation metrics. By controlling these experimental factors, the study focuses on the representation mechanism itself rather than differences caused by inconsistent experimental settings.

This study presents a systematic investigation of deep learning approaches for automated chili leaf disease recognition. The experimental design enables a consistent comparison of five representative architectures, namely DenseNet, MobileNet, EfficientNet, ResNet, and Vision Transformer, under

the same classification setting. In addition, a two-phase transfer learning procedure is employed to progressively adapt knowledge acquired from general-purpose image datasets to the specific characteristics of chili leaf diseases. The study further examines how convolutional architectures, which primarily construct hierarchical feature representations, differ from Vision Transformer models that capture long-range relationships through global self-attention when processing disease symptoms with diverse appearances and spatial distributions. By integrating these components within a controlled experimental setting, the research offers empirical insights into how architectural characteristics and transfer learning procedures jointly influence the performance of automated chili leaf disease classification.

Through this framework, the study provides controlled empirical evidence on the interaction between model architecture and transfer learning strategy for automated chili leaf disease classification.

2. Methods

The methodology described in this section establishes the experimental procedure for evaluating CNN and Vision Transformer (ViT) models in chili leaf disease classification. It covers the characteristics and preparation of the dataset, the procedures followed during model training and processing, as well as the evaluation protocol used to maintain consistency across the two architectural approaches. The training process is divided into two main stages, as depicted in Figure 1, frozen backbone training and fine-tuning.

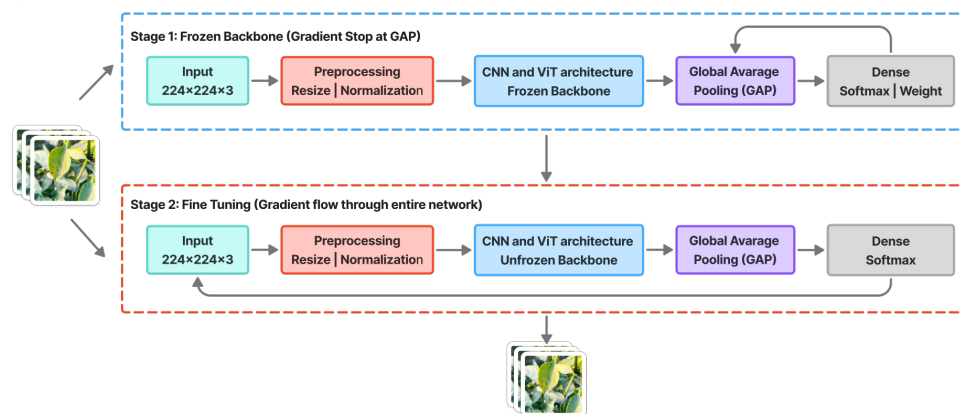


Figure 1. Proposed Methodology

2.1 Material

A total of 12,000 augmented chili leaf images were used in the experiments. The images originated from the Chilli Leaf Disease component of the Chilli Plant Leaf Disease and Growth Stage Dataset from Bangladesh, which is accessible through the Mendeley Data repository [27]. The selected data cover six classes, namely Bacterial Spot, Curl Virus, Cercospora Leaf Spot, Nutrition Deficiency, White Spot, and Healthy Leaves.

Rather than manually separating the images, the dataset was partitioned as part of the automated experimental pipeline using a directory-based generator. The predefined configuration assigned 70% of the available samples to model training, with the remaining 30% retained for validation and subsequent performance evaluation.

2.2 Implementation Details

The computational experiments were carried out using a personal computer featuring a 9th-generation Intel

Core i7 CPU, 64 GB DDR4 memory, and an NVIDIA GeForce GTX 1660 Ti graphics card with 6 GB of dedicated VRAM. The experimental environment was based on Windows 11, with Python 3.10 serving as the main programming language. TensorFlow and Keras were employed to develop and train the deep learning models, whereas Jupyter Notebook provided the environment for implementing and monitoring the experiments. GPU acceleration was enabled throughout the training process to reduce computational time, particularly for the Vision Transformer and CNN-based models.

All architectures were evaluated under identical computational and preprocessing conditions to maintain experimental consistency. The implementation followed a common pipeline covering image preprocessing, initialization from pretrained weights, the two-phase transfer learning procedure, training, validation, and final performance assessment. Keeping these factors constant reduces the influence of differences in hardware and software configurations, thereby making the comparison more reliable. Consequently, variations in the obtained results can be attributed primarily to differences in the model architectures and the adopted training strategy rather than to changes in the experimental environment.

2.3 Preprocessing

The preprocessing stage aimed to standardize the input dimensions and align the data distribution with the statistical characteristics anticipated by ImageNet-pretrained backbone models. This stage involved two primary operations: image resizing and intensity normalization.

Common preprocessing techniques in deep learning-based plant disease detection include resizing, normalization, and augmentation to standardize input images and improve model learning stability and generalization [28]. Resizing ensures that each image conforms to a fixed input dimension required by pretrained backbone architectures, which can significantly influence computational efficiency and classification performance. Let an input RGB image be represented by Equation 1.

$$I \in R^{H \times W \times 3} \quad (1)$$

H and W represent the original height and width of the image, respectively. To ensure compatibility with the CNN and Vision Transformer backbones employed in this study, each image is resized to a fixed spatial resolution of 224×224 . The resizing operation can be expressed as Equation 2:

$$I_r = \mathcal{R}(I), \quad I_r \in R^{224 \times 224 \times 3} \quad (2)$$

The coordinate mapping is given by Equation 3.

$$x = \frac{W}{224} x', \quad y = \frac{H}{224} y' \quad (3)$$

Since mapped coordinates are generally non-integer, bilinear interpolation is applied as in Equations 4 - 6.

$$x_0 = \lfloor x \rfloor, \quad x_1 = x_0 + 1,$$

$$y_0 = \lfloor y \rfloor, \quad y_1 = y_0 + 1 \quad (4)$$

$$d_x = x - x_0, \quad d_y = y - y_0 \quad (5)$$

$$\begin{aligned} I_r(x', y', c) &= (1 - d_x)(1 - d_y)I(x_0, y_0, c) \\ &+ d_x(1 - d_y)I(x_1, y_0, c) \\ &+ (1 - d_x)d_yI(x_0, y_1, c) + d_x d_y I(x_1, y_1, c) \end{aligned} \quad (6)$$

Normalization of pixel values following image resizing aligns the data distribution with that expected by pretrained networks, helping stabilize the gradient flow and improving convergence [29]. After resizing, pixel intensity values originally in the range $[0, 255]$ are first scaled to $[0, 1]$ is defined as Equation 7.

$$\tilde{I}(x, y, c) = \frac{I_r(x, y, c)}{255} \quad (7)$$

Image normalization is subsequently performed on each channel using the mean and standard deviation values established from ImageNet. The normalization is defined as Equation 8.

$$I_n(x, y, c) = \frac{\tilde{I}(x, y, c) - \mu_c}{\sigma_c} \quad (8)$$

Where the normalization parameters are defined as Equation 9 and 10.

$$\mu = [0.485, 0.456, 0.406] \quad (9)$$

$$\sigma = [0.229, 0.224, 0.225] \quad (10)$$

For each RGB channel, μ represents its mean value, whereas σ indicates the corresponding standard deviation, respectively. This normalization procedure serves three purposes: (i) to reduce distributional discrepancies between the chili leaf dataset and the ImageNet pretraining domain, (ii) to stabilize gradient propagation during optimization, and (iii) to facilitate faster convergence in the transfer learning process.

2.4 Convolutional Neural Network-Based Architecture

CNN are widely employed for image classification tasks due to their ability to learn hierarchical feature representations through convolutional operations [30], [31]. A CNN-based architecture typically consists of stacked convolutional layers, non-linear activation functions, normalization layers, and pooling operations that progressively transform raw pixel data into high-level semantic representations. Let the input image be defined as in Equation 11.

$$x \in R^{224 \times 224 \times 3} \quad (11)$$

The backbone CNN can be represented as a parametric function, as given by Equation 12.

$$h = f_{\theta}(x) \quad (12)$$

Where x denotes the convolutional feature extractor and h represents the final feature map. These feature maps are spatially aggregated using a Global Average Pooling (GAP) operation is given by Equation 13.

$$z = \text{GAP}(h) \quad (13)$$

The resulting feature vector x is then passed to a dropout layer and a fully connected classifier, producing the predicted probability distribution is presented by Equation 14.

$$y = \text{Softmax}(W_x + b) \quad (14)$$

CNN inherently possess a spatial inductive bias due to their localized receptive fields and parameter sharing. This allows shallow layers to extract low-level features like edges and textures. As the receptive field expands, deeper layers progressively encode more abstract and discriminative representations. Since chili leaf diseases are characterized by localized lesions, texture irregularities, and color deviations, CNN-based architectures are particularly suitable for capturing these subtle spatial variations. In this study, several CNN backbones are evaluated using a unified training protocol to ensure methodological consistency. Consequently, the reported performance differences primarily reflect architectural properties rather than optimization procedure disparities.

2.5 Vision Transformer Architecture

Unlike CNNs, Vision Transformers (ViT) rely on global self-attention mechanisms to model spatial dependencies [32]. Instead of applying convolutional filters, the input image is divided into non-overlapping patches, each of which is linearly embedded into a fixed-dimensional vector space [33]. Given an input image x , the number of patches is defined by Equation 15.

$$N = \frac{HW}{p^2} \quad (15)$$

H and W denote the height and width of the input image, respectively, and P denotes the patch size. Each patch embedding is processed by stacked transformer encoder layers composed of multi-head self-attention and feed-forward networks. The learnable class token z^{cls} serves as the global image representation, as given by Equation 16.

$$z^{cls} = f_\theta(x) \quad (16)$$

This representation is then passed to a classification head, like the one used in CNN-based models.

In contrast to CNNs, ViT does not rely on local receptive fields. Instead, it directly models long-range dependencies by computing attention weights between all patches [34]. This enables global contextual reasoning, which may be beneficial for capturing distributed disease patterns, such as large-scale discoloration or complex structural distortions across the leaf surface. In this study, a ViT-B16 backbone pretrained on ImageNet is employed. To ensure fair comparison, identical preprocessing, classifier design, and two-stage transfer learning procedures are applied across both CNN and ViT architectures.

2.6 Two-Stage Transfer Learning Framework

This study adopts a two-stage transfer learning strategy applied uniformly across all evaluated architectures, including DenseNet, ResNet, EfficientNet, MobileNet, and Vision Transformer [24-26]. The framework consists of a frozen training phase followed by controlled fine-tuning, ensuring stable adaptation from ImageNet-pretrained weights to the Chili leaf disease dataset. The complete model can be defined as Equation 17.

$$\hat{y} = g_\phi(f_\theta(x)) \quad (17)$$

x denotes the pretrained backbone, and $g_\phi(x)$ represents the classification head. In the first stage, all backbone parameters are frozen, as given by Equation 18.

$$\theta = \text{fixed} \quad (18)$$

The classification head is the only part optimized, as per Equation 19.

$$\min_{\phi} \mathcal{L}(y, g_\phi(f_\theta(x))) \quad (19)$$

This stage performs linear adaptation of the classifier while preserving the generic visual representations learned from ImageNet. The objective is to stabilize training and prevent abrupt distortion of pretrained features. For CNN architectures, the backbone outputs feature maps that are converted into fixed-length vectors using Global Average Pooling (GAP). For ViT, the class token embedding z^{cls} is directly used as the global representation.

In the second stage, the backbone is partially unfrozen to facilitate domain adaptation. The backbone parameters are decomposed as per Equation 20.

$$\theta = \{\theta_{early}, \theta_{late}\} \quad (20)$$

Early layers remain frozen as per Equation 21.

$$\theta_{early} = \text{fixed} \quad (21)$$

Only higher-level layers are updated according to Equation 22.

$$\min_{\theta_{late}, \phi} \mathcal{L}(y, g_\phi(f_\theta(x))) \quad (22)$$

During this phase, the learning rate is reduced, as described by Equation 23.

$$\eta_{fine} = 0.1 \quad (23)$$

This controlled fine-tuning process refines high-level semantic representations while preserving low-level generic features. Table 1 offers a comprehensive overview of the complete two-stage transfer learning configuration applied uniformly across CNN-based and Vision Transformer architectures. All models underwent identical preprocessing, optimizer settings, Learning Rate (LR) schedules, and callback strategies during training.

2.7 Evaluation Metrics

To ensure an unbiased evaluation of the multi-class Chilli leaf disease classification task, four standard

metrics are utilized. Their mathematical definitions are presented in Table 2. Here, *TP*, *TN*, *FP*, and *FN* represent true positives, true negatives, false positives, and false negatives, respectively.

Table 1. Two-Stage Transfer Learning Procedure

Component	CNN Architecture	ViT
Input Size	224x224x3	224x224x3
Preprocessing	Resize+Normalization	Resize+Normalization
Global Representation	Global Average Pooling	Class Token
Classifier Head	Dropout(0.3)+Dense(6)+Softmax	LayerNorm+Dropout(0.3)+Dense(6)+Softmax
Early Stopping Patience	3	3
ReduceLRonPlateau Factor	0.5	0.5
ReduceLRonPlateau Patience	2	2
Initial LR	1×10^{-4}	1×10^{-4}
Minimum LR	1×10^{-7}	1×10^{-7}
Optimizer	Adam	Adam
Loss	SparseCategoricalCrossentropy	SparseCategoricalCrossentropy
Metrics	Accuracy	Accuracy
Stage 1- Frozen Backbone training		
Backbone	Fully frozen	Fully frozen
Epoch	5	5
LR	1×10^{-4}	1×10^{-4}
Stage 2- Partial Fine Tuning		
Unfrozen layer	Last~30% conv blocks	Last~30% transformer blocks
Epoch	10	10
LR	1×10^{-5} (0.1 $\times$ initial LR)	1×10^{-5} (0.1 $\times$ initial LR)

Table 2. Performance Metrics

Metric	Formula
Accuracy (Acc)	$(TP+TN)/(TP+TN+FP+FN)$
Precision (P)	$TP/(TP+FP)$
Recall (R)	$TP/(TP+FN)$
F1-score	$2(PXR)/(P+R)$

For the multi-class classification task, both macro and weighted-average scores are reported. Model selection is based on the highest validation accuracy, with early stopping and learning rate scheduling implemented to ensure stable convergence. All reported metrics are computed on the held-out evaluation subset.

3. Results and Discussions

The performance obtained from the proposed two-stage transfer learning approach is examined in this section for chili leaf disease recognition. The experimental analysis covers four convolutional neural network models and one transformer-based model. Each architecture underwent the same training and evaluation procedure, including both the initial frozen-backbone phase and the subsequent fine-tuning phase. Applying an identical experimental protocol across all models enables a consistent assessment of their classification capabilities and facilitates a meaningful comparison between convolution-based and self-attention-based architectures.

3.1 Dynamics Training Analysis

The accuracy and loss trends observed during training and validation are illustrated in Figure 2 for the two-stage transfer learning approach. All architectures, including CNN-based models and ViT-B16, exhibit stable convergence. In Stage 1, the validation accuracy steadily increases while the loss decreases smoothly,

indicating effective adaptation of the classification head. In Stage 2, the loss declines sharply and the accuracy improves rapidly beyond 99%. The close alignment of the training and validation curves suggests strong generalization. Convergence stabilizes quickly, and early stopping prevents overfitting.

Table 3. Summary of Stage-wise Improvements

Architectures	Val Acc (Stage 1)	Val Acc (Stage 2)	Absolute Gain
DenseNet121	0.9475	0.9950	+0.0474
EfficientNetB0	0.9464	0.9950	+0.0486
MobileNetV2	0.9369	0.9961	+0.0592
ResNet152V2	0.9230	0.9939	+0.0709
ViT-B16	0.9810	0.9950	+0.0140

Table 3 summarizes the stage-wise validation accuracy across all evaluated models. All architectures show performance improvement in Stage 2, confirming the effectiveness of the proposed two-stage transfer learning strategy.

CNN-based models demonstrate more significant gains between stages, particularly ResNet152V2 and MobileNetV2. This behavior suggests that convolutional architectures benefit greatly from selective unfreezing, which allows for the refinement of higher-level feature maps to capture domain-specific patterns. During Stage 1, frozen backbones primarily function as generic feature extractors, while Stage 2 enables adaptation to disease-specific characteristics through controlled fine-tuning. In contrast, ViT-B16 achieves strong performance even in Stage 1, with only moderate improvement observed in Stage 2. This indicates that pretrained transformer representations are more readily transferable to the target domain, likely due to the global self-attention mechanism that captures holistic image context from the early stage.

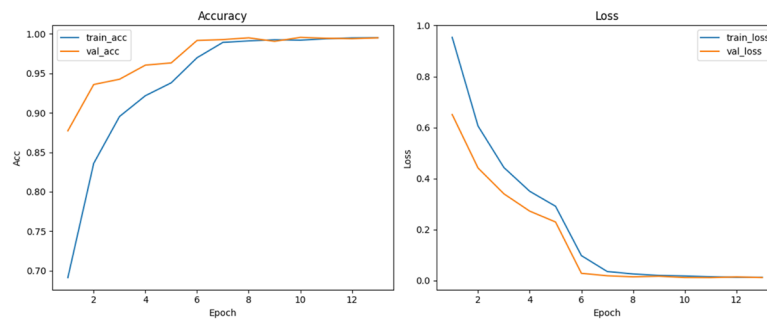
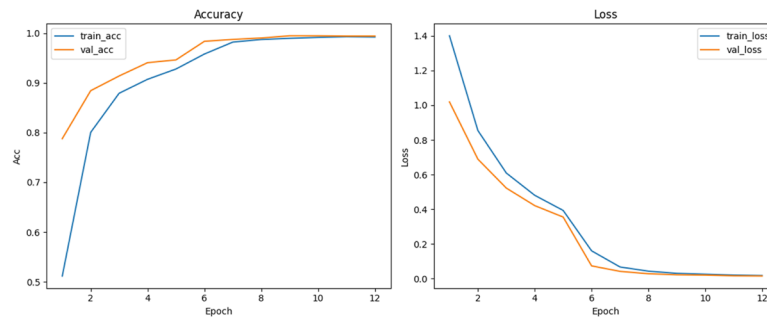
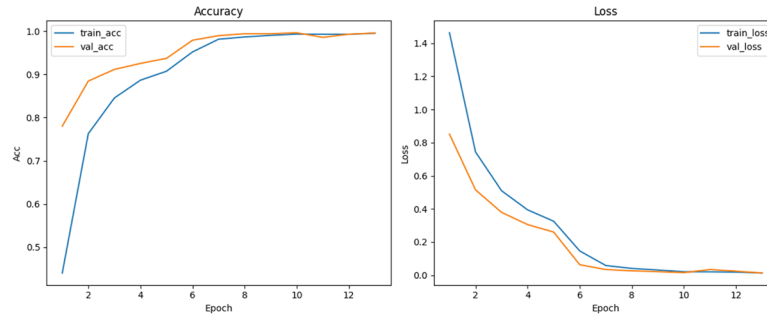
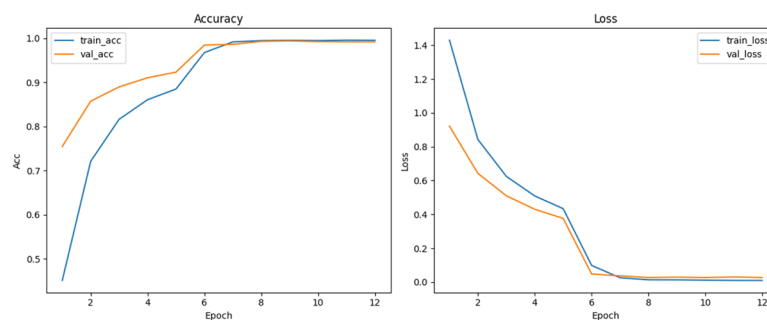
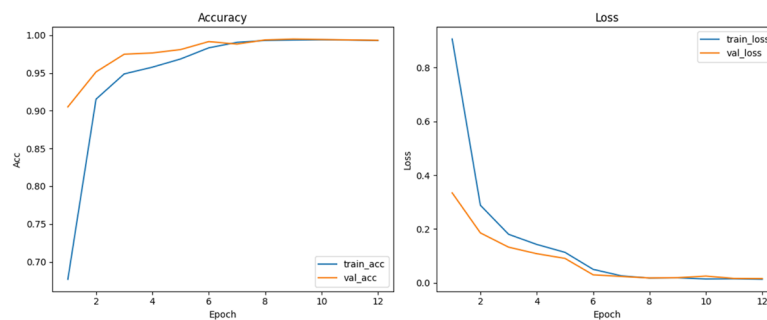
DenseNet121**EfficientNetB0****MobileNetV2****ResNet152V2****VITB16**

Figure 2. Accuracy and Loss Graph for each architecture

The observed differences highlight that the impact of the two-stage strategy varies depending on the architecture.

While CNN models rely more on fine-tuning to adapt localized features, Vision Transformer models benefit from strong initial representations and require less adjustment.

3.2 Testing Performance Evaluation

Generalization capability was assessed by using the model checkpoints that achieved the best validation performance for subsequent evaluation on an independent test set. The resulting classification performance is reported in Table 4 using four metrics: accuracy, macro-precision, macro-recall, and macro F1-score, calculated over the six disease categories.

The results show that every evaluated architecture obtained a test accuracy exceeding 99.50%, demonstrating that the models maintained consistently high predictive performance on previously unseen samples under the adopted training framework.

Table 4. Summary of Overall Performance Evaluation

Architectures	Acc	Macro Precision	Macro Recall	Macro F1 Score
DenseNet121	0.9956	0.9956	0.9955	0.9955
EfficientNetB0	0.9950	0.9951	0.9950	0.9955
MobileNetV2	0.9950	0.9950	0.9949	0.9949
ResNet152V2	0.9950	0.9951	0.9949	0.9950
ViT-B16	0.9961	0.9961	0.9961	0.9961

ViT-B16 achieves the highest accuracy of 99.61%, accompanied by consistent macro-averaged metrics, suggesting slightly better class-balanced predictions. DenseNet121 follows closely with an accuracy of 99.56%, while EfficientNetB0, MobileNetV2, and ResNet152V2 exhibit comparable performance with minimal variation.

The small differences observed among the evaluated models suggest that the training strategy has a greater influence on the final classification results than the choice of network architecture. By stabilizing feature extraction during the initial stage and refining higher-level representations during fine tuning, the framework enables effective adaptation across different backbone types.

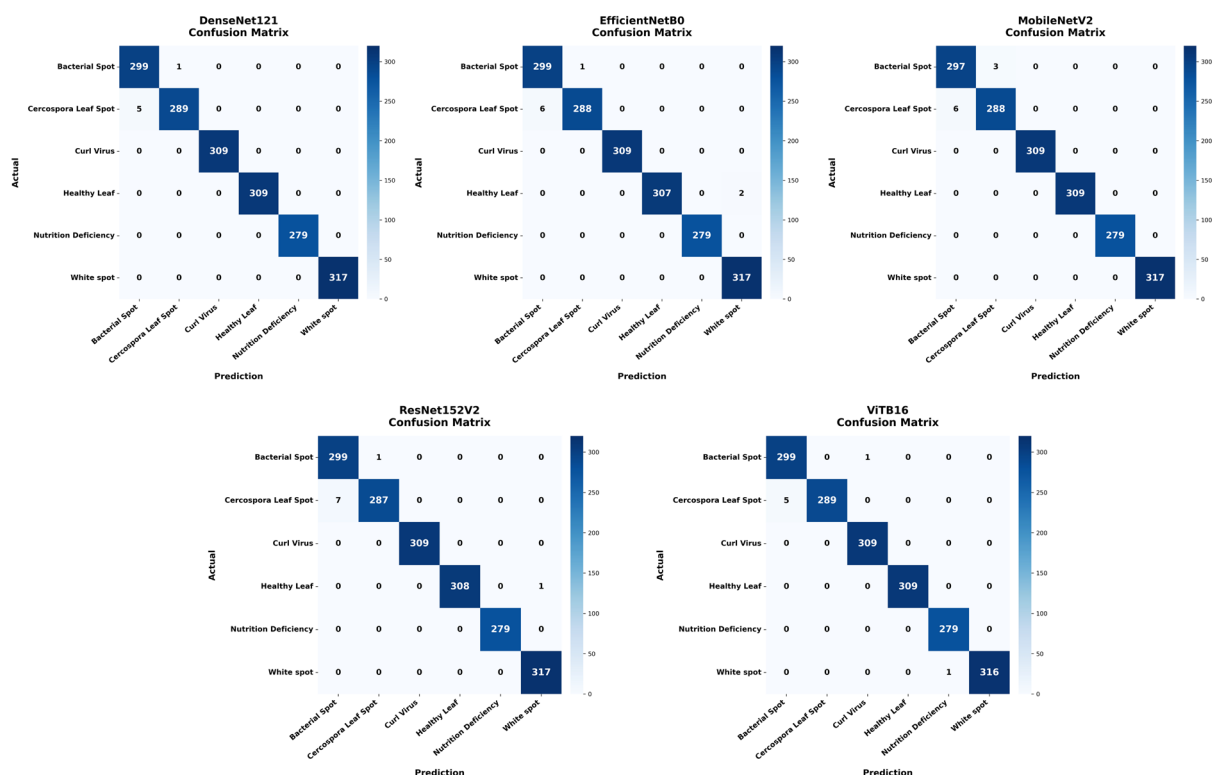


Figure 3. Confusion Matrix for each architecture

Furthermore, the slightly higher consistency observed in ViT-B16 suggests that global attention mechanisms may provide advantages in capturing distributed disease patterns, particularly for classes with subtle visual differences. However, the overall similarity in performance implies that, when combined with appropriate optimization scheduling, both convolutional and transformer-based models can achieve highly competitive results.

3.3 Confusion Matrix Analysis

The class-wise prediction patterns obtained on the test data are visualized as confusion matrices in Figure 3. Across the evaluated architectures, the predictions are concentrated predominantly along the main diagonal, demonstrating reliable discrimination among the disease categories. Only a small number of errors are

observed, mainly involving classes with visually similar symptoms.

Minor cross-predictions are primarily observed between Bacterial Spot and Cercospora Leaf Spot. This is likely because of overlapping morphological traits, such as spot-like lesions and similar color patterns. This suggests that the model has difficulty distinguishing between these classes within a narrow feature space, rather than being unstable. Architecturally, CNN-based models effectively capture local texture cues but may be limited in modeling globally distributed lesion patterns. In contrast, ViT-B16 demonstrates slightly improved separation of similar classes, indicating that global self-attention enhances contextual integration across the leaf surface.

3.4 RoC Curve Analysis

The multiclass discrimination performance of the evaluated models is illustrated through the one-vs-rest

ROC analysis in Figure 4. Most of the curves remain close to the ideal upper-left boundary, while the corresponding AUC scores are close to 1.000 for the majority of disease categories. This behavior indicates that the models can effectively distinguish among the different classes with minimal overlap in their prediction scores.

Despite differences in their underlying architectures, the CNN models and ViT-B16 exhibit highly comparable ROC profiles. This consistency suggests that the adopted two-stage training procedure can effectively adapt the pretrained models and establish well-defined class boundaries across different backbone designs. The ROC findings also agree with the results obtained from the accuracy measurements and confusion matrices, supporting the overall stability and generalization capability of the evaluated models.

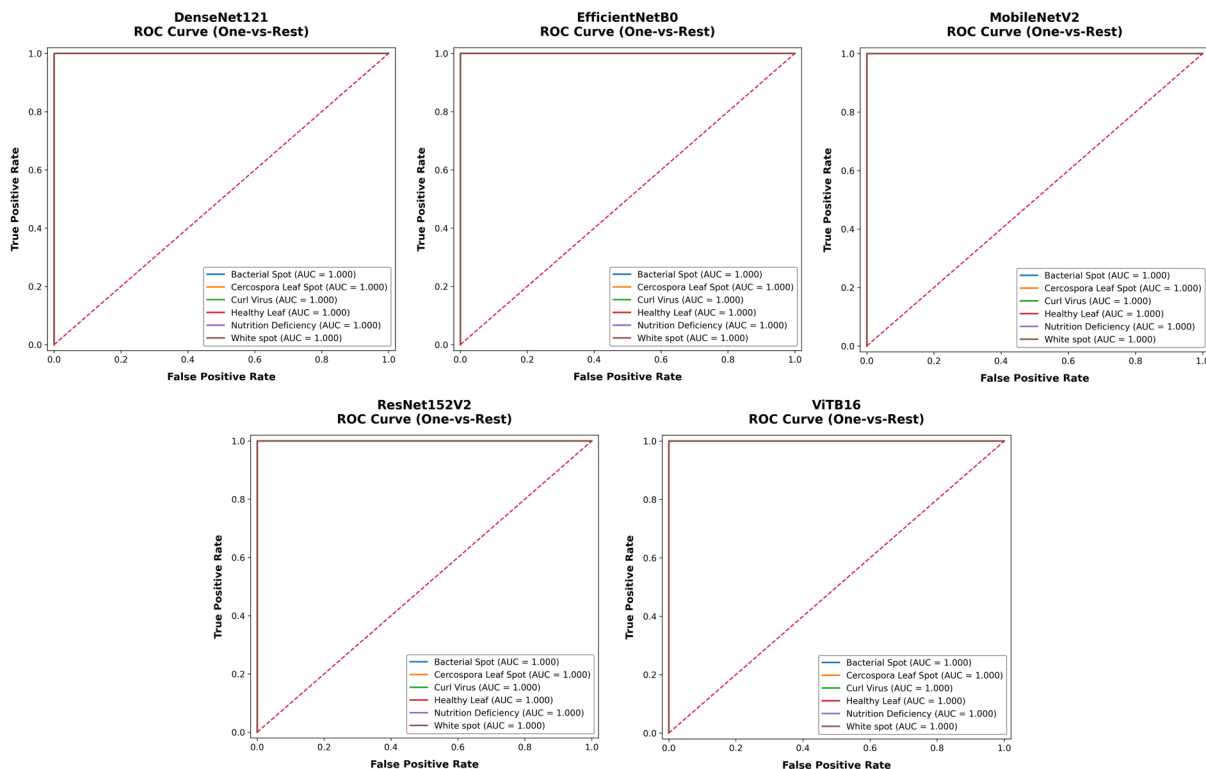


Figure 4. RoC Curve for each architecture

3.5 Discussions

A comparison between the results obtained in this study and those reported in earlier chili leaf disease classification research is provided in Table 5. The reported accuracy of previous methods ranges from 93.00% to 99.56%, with performance varying across CNN, ensemble, traditional machine learning, and Vision Transformer approaches. Among the previous studies, SEDCNN and RNDDNet achieved accuracies of 98.86% and 98.09%, respectively [10], [11], while an EfficientNet-based ensemble reported 99.56% [16]. The ViT-based approach achieved 99.35% accuracy [23].

Using the Mendeley dataset of 12,000 images representing six disease categories, ViT-B16 combined with the two-stage training procedure attained a test accuracy of 99.61%. This result is 0.26 percentage points higher than the previous ViT-based approach [23] and 0.05 percentage points higher than the EfficientNet-based ensemble [16]. These numerical differences, however, should be considered in the context of the experimental conditions rather than viewed as definitive evidence of model superiority, since the studies differ in dataset composition, class distribution, and evaluation protocols. In contrast to a direct comparison based solely on accuracy, the principal objective of this work is to assess several

architectures within a common experimental setting, using consistent preprocessing, data partitioning, and a two-stage transfer learning procedure.

Overall, the experimental approach achieved competitive performance for six-class chili leaf disease classification. The controlled experimental setting also provides a consistent basis for examining the relationship between model architecture and transfer learning strategy. Nevertheless, direct comparison across studies remains limited by differences in dataset characteristics and evaluation procedures.

3.6 Limitations and Future Works

Although the proposed models achieved high classification accuracy, several factors may constrain the broader applicability of the findings. The dataset, despite containing a relatively large number of images, was collected within a particular image acquisition environment. Consequently, model performance may not transfer directly to practical agricultural settings, where changes in illumination, background conditions, leaf occlusion, image quality, and camera characteristics can be substantially greater. Similar performance variability has been observed in previous studies that employed images captured under field conditions [17]. Future investigations should therefore validate the proposed framework on independent datasets and images collected directly in agricultural fields to determine its robustness and cross-domain generalization capability.

Second, although ViT-B16 delivered the best classification accuracy among the evaluated models, its relatively high computational demand may hinder practical implementation on resource-constrained devices, particularly when compared with lightweight architectures such as MobileNetV2. Future research may address this limitation by exploring techniques such as model compression, network pruning, parameter quantization, and knowledge distillation. These optimization strategies can reduce computational and memory overhead while preserving an acceptable level of predictive performance. Similar techniques have previously been investigated to facilitate the deployment of deep learning-based agricultural image classification systems in resource-limited environments [18].

Finally, the scope of this work is limited to determining the disease category from each input image, without estimating where the symptoms occur on the leaf or how severe the infection is. A potential direction for future research is to incorporate object detection and lesion localization techniques, together with severity assessment, so that the resulting system can deliver more comprehensive information for real-world agricultural decision-making [17]. Further investigation should also consider the computational demands and inference efficiency of the developed models when deployed on mobile and edge-computing platforms, providing a more practical assessment of their suitability for field applications.

Table 5. Comparison of proposed framework with related work on Chilli leaf disease classification

Ref.	Study Focus	Type	Architecture	Dataset used	Acc. (%)
[8]	CNN-based classification (4 classes)	CNN	InceptionV3	2,000 chili leaf images (4 classes)	93.00
[9]	Traditional ML-based classification	ML	Modified SVM	Standard plant disease database	95.00
[10]	SE-integrated Dense CNN	CNN	SEDCNN (SE-based Dense CNN)	Chili leaf dataset (6 classes)	98.86
[11]	Residual nested dilated DenseNet	CNN	RNDDNet	3800 Cilli leaf image (6 classes)	98.09
[12]	Ensemble deep learning framework	Ens. CNN	Chilli-Net	Chilli leaf disease	97.00
[13]	Lightweight CNN architecture	CNN	Modified MobileNet	Chilli leaf disease	97.18
[14]	CNN with augmentation	CNN	Standard CNN	Indonesia Chilli dataset (3 classes)	95.89
[15]	Pretrained CNN feature extraction	CNN	EfficientNet-B0	Chilli leaf disease	97.06
[16]	Ensemble EfficientNet variants	Ens. CNN	EfficientNet-based ensemble	BARI-approved Chilli dataset	99.56
[23]	Transformer-based classification	ViT	ViT-Based model	Proprietary dataset (6 classes)	99.35
Proposed	Two-stage transfer learning+ViT	Transfer learning, ViT	ViT-B16	Mendeley dataset (12,000 images, 6 classes)	99.61

4. Conclusions

This study presented a structured two-stage transfer learning framework for chili leaf disease classification and evaluated its performance across multiple convolutional neural network architectures and a Vision Transformer. Under the experimental conditions used in this study, the proposed strategy improved

classification performance across the evaluated architectures, with ViT-B16 achieving the highest testing accuracy of 99.61%. The results indicate that the combination of controlled backbone freezing and subsequent selective fine-tuning can effectively adapt pretrained models to the chili leaf disease classification task.

The comparative experiments also show that transfer learning strategy should be considered alongside model architecture when developing image-based plant disease classification systems. The consistent evaluation protocol applied across the architectures provides a controlled basis for comparing their performance, while the use of a publicly available dataset facilitates reproducibility.

Several directions remain for further investigation. Future work should include cross-dataset and field-based validation, evaluation on resource-constrained devices, and extension from image-level classification toward disease localization and severity estimation. These developments would help assess the applicability of the proposed framework under more diverse and practical agricultural conditions.

Declarations

AI Usage Statement

The authors used an AI-assisted language tool solely to improve the readability, grammar, and language clarity of the manuscript. The AI tool was not used to generate, analyze, or interpret research findings. All scientific content, experimental procedures, data analysis, interpretations, and conclusions were developed, verified, and approved by the authors.

Author Contributions

R.K.: conceptualization, methodology, software, data curation, formal analysis, investigation, writing – original draft, and visualization. S.S.: methodology, formal analysis, validation, and writing – review and editing. F.S.M.: methodology, validation, and writing – review and editing. B.I.: data curation, formal analysis, visualization, and writing – review and editing. N.K.D.: supervision, validation, and writing – review and editing. All authors reviewed and approved the final version of the manuscript.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The research was conducted without external financial support.

Conflict of Interest Statement

The authors declare no conflict of interest.

Ethical Clearance

This study used a publicly available image dataset and did not involve human participants, personally identifiable information, or animal subjects. Therefore, ethical clearance was not required for this study.

Data and Code Availability

The dataset used in this study is publicly available through the Mendeley Data repository: <https://data.mendeley.com/datasets/w9mr3vf56s/1>. The dataset is identified as Mendeley Data, DOI:

10.17632/w9mr3vf56s.1. The data can be accessed by other researchers to support the reproducibility and further evaluation of the reported experiments. The source code developed for the experiments is not publicly available at the time of publication but can be made available by the corresponding author upon reasonable request.

References

- [1] V. K. Velpula, S. Prasad, J. Vadlamudi, S. Gurralla, A. V. Priya, and K. Srinidhi, "Chilli Leaf Disease Prediction System using Deep Learning Model," in *2025 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, Mar. 2025, pp. 1–6. doi: [10.1109/ESCI63694.2025.10988404](https://doi.org/10.1109/ESCI63694.2025.10988404).
- [2] N. Krishnan et al., "Prevalence of chilli leaf curl virus and tomato leaf curl New Delhi virus with chilli leaf curl disease in India," *Physiology and Molecular Biology of Plants*, vol. 31, no. 8, pp. 1363–1374, Aug. 2025, doi: [10.1007/s12298-025-01570-w](https://doi.org/10.1007/s12298-025-01570-w).
- [3] S. Saha and K. Tayung, "First report of Fusarium proliferatum causing leaf blight of chilli (*Capsicum annuum*) in India," *Crop Protection*, vol. 190, p. 107094, Apr. 2025, doi: [10.1016/j.cropro.2024.107094](https://doi.org/10.1016/j.cropro.2024.107094).
- [4] B. N. Naik, R. Malmathanraj, and P. Palanisamy, "Detection and classification of chilli leaf disease using a squeeze-and-excitation-based CNN model," *Ecol. Inform.*, vol. 69, p. 101663, Jul. 2022, doi: [10.1016/j.ecoinf.2022.101663](https://doi.org/10.1016/j.ecoinf.2022.101663).
- [5] S. A. Amjad, T. Anuradha, T. M. Datta, and U. M. Babu, "Chilli Leaf Disease Detection Using Deep Learning," 2024, pp. 81–89. doi: [10.1007/978-3-031-56703-2_7](https://doi.org/10.1007/978-3-031-56703-2_7).
- [6] V. K. Pratap and N. S. Kumar, "High-precision multiclass classification of chili leaf disease through customized EfficientNetB4 from chili leaf images," *Smart Agricultural Technology*, vol. 5, p. 100295, Oct. 2023, doi: [10.1016/j.atech.2023.100295](https://doi.org/10.1016/j.atech.2023.100295).
- [7] W. S. Hettiarachchi and H. M. D. S. Herath, "AI-based Chilli Leaf Disease Detection and Remedy Recommendation System," in *2025 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAET)*, IEEE, Aug. 2025, pp. 449–454. doi: [10.1109/IICAET67254.2025.11265591](https://doi.org/10.1109/IICAET67254.2025.11265591).
- [8] Z. Gulzar, S. Chandu, and K. Ravi, "Classification and Analysis of Chilli Plant Disease Detection Using Convolution Neural Networks," 2023, pp. 677–696. doi: [10.1007/978-981-99-7093-3_45](https://doi.org/10.1007/978-981-99-7093-3_45).
- [9] R. Thyagaraj., T. Y. Satheesha, and S. Bhairannawar, "Plant Leaf Disease Classification Using Modified SVM With Post Processing Techniques," in *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, 2023, pp. 1–4. doi: [10.1109/ICAISC58445.2023.10201001](https://doi.org/10.1109/ICAISC58445.2023.10201001).
- [10] N. V., Y. G., N. N. B., M. R., and P. P., "Empirical Analysis of Squeeze and Excitation-Based Densely Connected CNN for Chilli Leaf Disease Identification," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 4, pp. 1681–1692, Apr. 2024, doi: [10.1109/TAI.2024.3364126](https://doi.org/10.1109/TAI.2024.3364126).
- [11] M. Srinivasulu and S. Maiti, "RNDDNet: A residual nested dilated DenseNet based deep-learning model for chilli plant disease classification," *Engineering Research Express*, vol. 6, no. 3, p. 035204, Sep. 2024, doi: [10.1088/2631-8695/ad5f03](https://doi.org/10.1088/2631-8695/ad5f03).
- [12] S. Bandopadhyay, A. K. Gula, I. Haider, and G. Kumar, "Chili-Net: An Approach for Classifying Chili Leaf Diseases Using Deep Neural Networks," 2024, pp. 45–55. doi: [10.1007/978-3-031-66594-3_5](https://doi.org/10.1007/978-3-031-66594-3_5).
- [13] S. A. Hamim and A. I. Jony, "Enhanced deep learning model architecture for plant disease detection in Chilli plants," *Journal of Edge Computing*, vol. 3, no. 2, pp. 136–146, Nov. 2024, doi: [10.55056/jec.758](https://doi.org/10.55056/jec.758).
- [14] N. Lokeswari, D. Pramesti, and H. Fakhruroja, "Enhancing the Accuracy of Chili Plant Disease Classification using Convolutional Neural Networks and Image Augmentation," in *2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, IEEE,

- Jul. 2025, pp. 123–128. doi: [10.1109/IAICT65714.2025.11100595](https://doi.org/10.1109/IAICT65714.2025.11100595).
- [15] N. F. Aminuddin, H. Abdul Kadir, M. R. Md Tomari, A. Joret, and Z. Tukiran, "Towards Improved Disease Identification With Pretrained Convolutional Neural Networks As Feature Extractors For Chili Leaf Images," *J. Teknol.*, vol. 86, no. 2, pp. 89–100, Jan. 2024, doi: [10.11113/jurnalteknologi.v86.19853](https://doi.org/10.11113/jurnalteknologi.v86.19853).
- [16] I. I. Adnan, M. Z. Waughfa, and T. Ahmed, "Enhancing Chili Plant Disease Detection: A Comprehensive Ensemble Approach Utilizing Pre-Trained Deep Learning Models," in *2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEELACON)*, IEEE, Sep. 2024, pp. 1–6. doi: [10.1109/PEELACON63629.2024.10800382](https://doi.org/10.1109/PEELACON63629.2024.10800382).
- [17] M. K. M. A. Ranjan, and R. Machavaram, "In-field Chilli Crop Disease Detection Using YOLOv5 Deep Learning Technique," in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2023, pp. 1–6. doi: [10.1109/I2CT57861.2023.10126468](https://doi.org/10.1109/I2CT57861.2023.10126468).
- [18] S. Rahman et al., "Reducing Computational Cost in Chili Leaf Disease Classification Using Optimized RegNet," *IAENG Int. J. Comput. Sci.*, vol. 52, no. 12, pp. 4940–4951, Dec. 2025, Accessed: Sep. 20, 2026. [Online]. Available: https://www.iaeng.org/IJCS/issues_v52/issue_12/IJCS_52_12_36.pdf
- [19] T. R. G. G. Devadhas, and T. Y. Satheesha, "EGMEO: Exponential Groupers and Moray Eels Optimization Based Googlenet for Chili Leaf Disease Classification," in *2025 International Conference on Next Generation Computing Systems (ICNGCS)*, IEEE, Aug. 2025, pp. 1–7. doi: [10.1109/ICNGCS64900.2025.11183425](https://doi.org/10.1109/ICNGCS64900.2025.11183425).
- [20] M. Ahmad, T. M. A. Zulcaffle, and M. A. Khan, "Advance Agricultural Precision: Chili Plant Disease Prediction by DenseNet121 with images," in *2025 IEEE International Conference on Emerging Trends in Engineering and Computing (ETECOM)*, IEEE, Oct. 2025, pp. 1–7. doi: [10.1109/ETECOM66111.2025.11319030](https://doi.org/10.1109/ETECOM66111.2025.11319030).
- [21] S. Mehdipour, S. A. Mirroshandel, and S. A. Tabatabaei, "Vision transformers in precision agriculture: A comprehensive survey," *Intelligent Systems with Applications*, vol. 29, p. 200617, Mar. 2026, doi: [10.1016/j.iswa.2025.200617](https://doi.org/10.1016/j.iswa.2025.200617).
- [22] S. Aboelenin, F. A. Elbasheer, M. M. Eltoukhy, W. M. El-Hady, and K. M. Hosny, "A hybrid Framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer," *Complex & Intelligent Systems*, vol. 11, no. 2, p. 142, Feb. 2025, doi: [10.1007/s40747-024-01764-x](https://doi.org/10.1007/s40747-024-01764-x).
- [23] T. K. V. B. C. and M. S. Srividya, "Detection of Chilli Leaf Diseases Using Vision Transformers and Attention Mechanisms," in *2025 9th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)*, IEEE, Nov. 2025, pp. 1–5. doi: [10.1109/CSITSS67709.2025.11294055](https://doi.org/10.1109/CSITSS67709.2025.11294055).
- [24] T. Sanida, A. Sideris, M. V. Sanida, and M. Dasygenis, "Tomato leaf disease identification via two-stage transfer learning approach," *Smart Agricultural Technology*, vol. 5, p. 100275, Oct. 2023, doi: [10.1016/j.atech.2023.100275](https://doi.org/10.1016/j.atech.2023.100275).
- [25] J. Meng, Z. Tan, Y. Yu, P. Wang, and S. Liu, "TL-med: A Two-stage transfer learning recognition model for medical images of COVID-19," *Biocybern. Biomed. Eng.*, vol. 42, no. 3, pp. 842–855, Jul. 2022, doi: [10.1016/j.bbe.2022.04.005](https://doi.org/10.1016/j.bbe.2022.04.005).
- [26] X. Zhang, Z. Chen, J. Gao, W. Huang, P. Li, and J. Zhang, "A two-stage deep transfer learning model and its application for medical image processing in Traditional Chinese Medicine," *Knowl. Based. Syst.*, vol. 239, p. 108060, Mar. 2022, doi: [10.1016/j.knsys.2021.108060](https://doi.org/10.1016/j.knsys.2021.108060).
- [27] M. A. S. Nirob, A. K. M. F. K. Siam, P. Bishshash, and M. Assaduzzaman, "Chili Plant Leaf Disease and Growth Stage Dataset from Bangladesh," 2025, *Mendeley Data*. doi: [10.17632/w9mr3vf56s.1](https://doi.org/10.17632/w9mr3vf56s.1).
- [28] N. Paul, G. C. Sunil, D. Horvath, and X. Sun, "Deep learning for plant stress detection: A comprehensive review of technologies, challenges, and future directions," *Comput. Electron. Agric.*, vol. 229, p. 109734, Feb. 2025, doi: [10.1016/j.compag.2024.109734](https://doi.org/10.1016/j.compag.2024.109734).
- [29] K. Hong, Y. Zhou, and H. Han, "The pipelines of deep learning-based plant image processing," *Quantitative Plant Biology*, vol. 6, p. e23, Jul. 2025, doi: [10.1017/qpb.2025.10018](https://doi.org/10.1017/qpb.2025.10018).
- [30] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017, doi: [10.1016/j.neucom.2016.12.038](https://doi.org/10.1016/j.neucom.2016.12.038).
- [31] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, Jun. 2020, doi: [10.1007/s13748-019-00203-0](https://doi.org/10.1007/s13748-019-00203-0).
- [32] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ArXiv*, 2020. doi: [10.48550/arXiv.2010.11929](https://doi.org/10.48550/arXiv.2010.11929).
- [33] K. T. Chitty-Venkata, S. Mittal, M. Emani, V. Vishwanath, and A. K. Somani, "A survey of techniques for optimizing transformer inference," *Journal of Systems Architecture*, vol. 144, p. 102990, Nov. 2023, doi: [10.1016/j.sysarc.2023.102990](https://doi.org/10.1016/j.sysarc.2023.102990).
- [34] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: [10.1145/3505244](https://doi.org/10.1145/3505244).