Journal of Computer Science

Research Article

A Novel Deep Learning Approach for Tomato Leaf Disease Detection Using Optimized CNN Architecture

Sunil K. Vithalani and Vipul K. Dabhi

Department of Information Technology, Dharmsinh Desai University, Nadiad, Gujarat, India

Article history
Received: 18-05-2025
Revised: 30-09-2025
Accepted: 05-08-2025

Corresponding Author:
Sunil K. Vithalani
Department of Information Technology, Dharmsinh Desai University, Nadiad, Gujarat, India
Email: sunilvithlani.it@ddu.ac.in

Abstract: Early disease detection in plants is essential for sustaining agricultural yield and guaranteeing food security. A comparative analysis of transformer-based and convolutional based deep learning models for classifying tomato leaf diseases is presented. Specifically, it examines the performance of a Vision Transformer (ViT), tested both in a form of scratch training setup and through transfer learning, against well-known CNN architectures such as Inception V3, VGG16, ResNet50, and a custom-designed lightweight CNN. This is one of the few studies to rigorously benchmark ViT against CNNs in the context of agricultural disease detection using the PlantVillage dataset. The fine-tuned ViT model delivered the best results, achieving an accuracy of 95.53%, significantly outperforming all CNN counterparts. The lightweight CNN demonstrated strong performance with 93.12% accuracy, while offering clear benefits in terms of smaller model size and reduced computational cost making it well-suited for on-device or edge-level applications. Conversely, the ViT model trained from scratch underperformed due to dataset constraints, reinforcing the necessity of transfer learning for transformer architectures. Evaluation metrics included recall, accuracy, F1-score, and precision, which collectively illustrated the trade-off between high-capacity models and deployment feasibility. The main contribution of this work lies in introducing transformer-based learning into the plant pathology domain and the presentation of a scalable, low-computation alternative via lightweight CNNs. Future directions involve enlarging the dataset, integrating explainable AI techniques, and enabling real-time applications for precision agriculture.

Keywords: Convolutional Neural Network (CNN), Deep Learning (DL), Machine Learning, Plant Disease Classification, Tomato Leaf Disease Classification (TLDC), Transfer Learning

Introduction

Tomatoes (*Solanum lycopersicum*) are among the most widely cultivated horticultural crops globally, playing a critical role in food security and agricultural economies. According to Food and Agriculture Organization (FAO) statistics, global production surpasses 180 million tonnes annually (FAOSTAT, 2025), reflecting its importance in dietary consumption, processed foods, and industrial applications. Beyond their nutritional benefits, tomatoes are an essential income source for millions of smallholder and commercial farmers across diverse agro-climatic zones (Schreinemachers *et al*., 2018). Despite their significance, tomato crops are highly vulnerable to a wide array of bacterial, fungal, and viral diseases, which severely reduce yield quality and quantity (Manjunatha *et al*., 2025).

Among the most damaging pathogens are Septoria Leaf Spot, Late Blight, Early Blight, and Tomato Mosaic Virus each posing distinct diagnostic and control challenges (Alzahrani and Alsaade, 2023). Improper or delayed identification often results in the indiscriminate application of chemical treatments, such as fungicides and insecticides, which contribute to rising production costs, environmental degradation, and the emergence of pesticide-resistant pathogen strains (Sakkarvarthi *et al*., 2022). These conventional approaches not only compromise food safety but also hinder sustainable farming practices.

Conventional disease monitoring predominantly relies heavily on human visual inspection, which is labor-intensive, subjective, and prone to error especially in early stages where symptoms may be subtle or overlap across multiple diseases (Prajapati *et al*., 2017). Moreover, manual methods lack scalability, making them unsuitable for large-scale agricultural operations. These limitations highlight the urgent requirement for automated, accurate, and field-deployable disease detection systems.

The emergence of Artificial Intelligence (AI) and, in particular, DL has revolutionized the domain of automated plant disease detection. Convolutional Neural Networks (CNNs) achieved remarkable success in computer vision by learning hierarchical representations directly from raw image data (Li *et al*., 2021). While CNNs such as VGG16 (Simonyan and Zisserman, 2015) , Inception V3 (Szegedy *et al*., 2016), and ResNet50 (He *et al*., 2016) have demonstrated strong classification performance but are computationally intensive due to their depth and parameter complexity.

To address this, our study introduces a comparative analysis involving:

1. Fine-tuned CNN models: VGG16, Inception V3, and ResNet50
2. A custom lightweight CNN with significantly fewer parameters
3. A transformer-based approach using the ViT-Base model (Dosovitskiy *et al*., 2021)

Unlike CNNs, the ViT divides an image into fixed-size patches and processes them using transformer encoders, enabling the capture of long-range dependencies and global contextual features often missed by conventional convolutional filters.

Literature Review

Crop productivity is strongly influenced by plant diseases, which can be broadly categorized as biotic (caused by pathogens such as fungi, bacteria, and viruses) or abiotic (resulting from environmental stresses like nutrient deficiencies, drought, or pollutants). Early and reliable detection is therefore critical for minimizing yield losses and ensuring sustainable production (Li *et al*., 2021).

DL approaches for plant disease recognition typically fall into three main categories: image classification, object detection, and semantic segmentation. Classification-based models assign entire images to predefined categories and are widely applied for disease presence/absence detection. CNN-based architectures such as Inception V3, and VGG16 have been particularly effective in extracting discriminative features from leaf images (Li *et al*., 2021). Object detection-based techniques not only classify but also localize affected areas using bounding boxes. While two-stage detectors such as R-CNN (Girshick *et al*., 2014) and Faster R-CNN (Ren *et al*., 2017) offer high accuracy, they are computationally expensive (Du *et al*., 2020). In contrast, one-stage detectors like YOLO (Redmon *et al*., 2016) and SSD (Liu *et al*., 2016) provide real-time detection capabilities at the cost of reduced accuracy. Segmentation-based approaches, which highlight the diseased regions at the pixel level, have been explored using DL models such as FCN (Long *et al*., 2015) and Mask R-CNN (Bondre and Patil, 2024), allowing precise identification of affected areas within the plant structure.

Several studies have investigated different CNN and other DL models for crop disease identification. Amara *et al*. (2017) employed the LeNet-CNN model to classify banana leaf diseases using a dataset from PlantVillage, achieving an accuracy of 99.72% by training the model on 3,700 resized images. Similarly, (Guerrero-Ibañez and Reyes-Muñoz, 2023) proposed a four-layer CNN for tomato leaf disease recognition, achieving 99.64% accuracy on a dataset combining Plant Village and field images, enhanced through GAN-based augmentation and K-fold cross-validation. (Trivedi *et al*., 2021) demonstrated an accuracy of 98.49% using an eight-layer deep neural network fine-tuned on the PlantVillage dataset for early tomato disease detection. Wang *et al*. (2017) explored DL for apple leaf disease severity classification and trained VGG19, VGG16, ResNet50, and Inception V3, models with transfer learning, achieving 90.4% accuracy with VGG16 on a dataset containing 17,640 images. Kerkech *et al*. (2020) introduced an innovative approach using UAV-based vineyard images captured with RGB and infrared cameras. They segmented the images using separate SegNet models for each modality and achieved a classification accuracy of 95.02% by implementing a fusion strategy that combined RGB and infrared data.

Karthik *et al*. (2020) investigated the effectiveness of traditional CNNs, residual CNNs, and attention-based residual CNNs for tomato leaf disease detection using an extensive dataset from Plant Village. Their dataset, augmented to 95,999 images, was validated using a five-fold cross-validation strategy, resulting in an accuracy of 98% for the attention-based residual CNN model, which effectively focused on disease-relevant image features. Picon *et al*. (2019), explored both single and multi-crop classification models using images captured in real field conditions via mobile devices. Their dataset consisted of 121,955 images spanning several crops, including wheat, rapeseed, corn, rice, and barley. By employing ResNet50, they compared independent crop-specific models against a multi-crop classification model and found that the latter slightly outperformed single-crop models, achieving an

accuracy of 98%. Their results further suggested that incorporating crop-type information during training improved model generalization across different plant species and environmental conditions.

While these studies demonstrate significant advancements in plant disease classification, several research gaps persist. A primary limitation is the lack of generalization across real-world agricultural environments. Many models are trained on controlled datasets like Plant Village (Hughes and Salathe, 2015), which contain high-resolution, controlled images. However, these models often struggle to maintain high accuracy when deployed in real-world conditions, where lighting variations, occlusions, and background noise introduce significant challenges. Multi-disease detection remains underexplored, as most studies assume only one disease per sample, despite the common occurrence of co-infections (Kamilaris and Prenafeta-Boldú, 2018).

Furthermore, most high-accuracy models rely on computationally expensive architectures, making them unsuitable for distribution on low-power devices such as UAVs and mobile phones. The development of lightweight, optimized models that maintain high performance while operating on low resource devices is essential for practical applications in precision agriculture (Peyal *et al*., 2023). The lack of transparency in DL models creates significant challenges for their interpretability and trust. To support informed decisions in disease management and pesticide use, farmers and agronomists need explainable AI methods that clarify model predictions (Samek *et al*., 2017). Moreover, while some studies have incorporated RGB and infrared imaging (Kerkech *et al*., 2020), the potential of multispectral and hyperspectral imaging remains underutilized. These advanced imaging modalities could enhance disease differentiation by capturing spectral signatures beyond the visible range, enabling more precise identification of disease symptoms (Arnal Barbedo, 2019).

Another challenge in plant disease classification is the dependency on large amounts of labelled training data. Many existing models require extensive datasets, which may not be available for all crops or disease types. Transfer learning and data augmentation techniques, such as GAN-based synthetic image generation (Guerrero-Ibañez and Reyes-Muñoz, 2023), can be leveraged to address this limitation. Future research should explore cross-domain adaptation strategies that allow models trained on well-studied crops like tomatoes to be fine-tuned for less-documented crops with limited available data.

Although DL has greatly enhanced accuracy and efficiency of plant disease classification, challenges related to generalization, multi-disease detection, computational efficiency, interpretability, and data availability remain unresolved. Overcoming these challenges is crucial to realizing the practical use of DL models in precision agriculture, thereby enhancing crop health monitoring and promoting sustainable farming practices.

Comparative Analysis

Table 1 summarizes key studies on plant disease detection with deep learning.

Table 1: Comparison of Deep Learning Approaches for Plant Disease Detection

Study	Model Used	Dataset	Task Type	Accuracy (%)	Key Findings	Limitations
(Amara *et al*., 2017)	LeNet-CNN	PlantVillage (Banana leaves)	Classification	99.72	Achieved high accuracy for banana leaf diseases	Limited dataset, lacks real-field validation
(Guerrero-Ibañez and Reyes-Muñoz, 2023)	4-layer CNN + GAN-based augmentation	PlantVillage + Field Images (Tomato)	Classification	99.64	Combined real-world and synthetic data, preventing overfitting	Small model size but high accuracy; lacks multi-class detection
(Trivedi *et al*., 2021)	8-layer DNN	PlantVillage (Tomato)	Classification	98.49	Early tomato disease detection with deep learning	Does not generalize well to real-world conditions
(Wang *et al*., 2017)	VGG16, VGG19, Inception V3, ResNet50	PlantVillage (Apple leaves)	Classification	90.4 (VGG16)	Fine-tuned transfer learning models for apple leaf disease	High dependency on dataset, lacks multi-disease classification
(Kerkech *et al*., 2020)	SegNet (RGB + Infrared)	UAV-based Vineyard Dataset	Segmentation	95.02	Utilized RGB and infrared fusion for improved detection	Computationally expensive for real-time field deployment
(Karthik *et al*., 2020)	CNN, Residual CNN, Attention-Based CNN	PlantVillage (Tomato)	Classification	98	Attention mechanism improved feature selection	High training time and data augmentation required
(Picon *et al*., 2019)	ResNet50	Multi-Crop Dataset (121,955 images)	Classification	98	Multi-crop model improved generalization	Requires crop-type information for best performance

This comparative study highlights that while CNN-based approaches have demonstrated high classification accuracy, they often suffer from poor generalization in real-world conditions. Hybrid approaches that integrate multiple imaging modalities (e.g., RGB and infrared) show potential for improved performance but remain computationally expensive. Attention mechanisms and generative models such as GANs offer promising solutions to data scarcity but require extensive computational resources. Addressing these research gaps will be crucial in advancing DL applications for robust and scalable plant disease detection.

Research Gaps

Despite the developments in deep learning-based plant disease identification, several critical research gaps remain unaddressed:

- Limited Generalization to Real-World Conditions: Most existing models are trained on controlled datasets such as PlantVillage, which contain uniform lighting and background conditions. These models often struggle when deployed in real-field conditions with varying illumination, occlusions, and environmental noise (Ahmad *et al*., 2023; Picon *et al*., 2019)
- Lack of Multi-Disease Classification Models: Most existing studies concentrate on identifying a single disease in each image, while in real-world agricultural settings, multiple diseases may simultaneously affect the same plant. Developing multi-label classification models remains a challenge (Demilie, 2024)
- High Computational Costs and Resource Constraints: Numerous DL architectures demand substantial computational resources, limiting their suitability for mobile or edge device deployment. There is a need for lightweight and efficient models optimized for real-time disease detection in low-resource environments (Peyal *et al*., 2023)
- Limited Interpretability and Explainability: DL models function as "black boxes," making it difficult for farmers and agricultural experts to trust and understand predictions. Explainable AI (XAI) techniques must be integrated to improve model transparency (Samek *et al*., 2017)
- Underutilization of Advanced Imaging Modalities: While some studies incorporate RGB and infrared imaging, the potential of hyperspectral and multispectral imaging for precise disease differentiation remains underexplored (Kerkech *et al*., 2020)
- Dependence on Large Labeled Datasets: Many state-of-the-art models rely on large labeled datasets, which are often unavailable for certain crops and diseases. Transfer learning, synthetic data generation, and domain adaptation techniques need further exploration to mitigate this issue (Guerrero-Ibañez and Reyes-Muñoz, 2023)
- Inadequate Studies on Disease Progression and Severity Estimation: Most existing studies focus on disease presence or absence without assessing disease severity levels. Incorporating severity estimation models would allow farmers to take timely preventive measures (Wang *et al*., 2017)

Methods

This section outlines the dataset and model development strategies employed for TLDC. The PlantVillage dataset served as the benchmark for assessing DL models, with transfer learning employed to fine-tune pre-trained CNN architectures including Inception V3, ResNet50, and VGG16. A custom lightweight Convolutional Neural Network (CNN) was also designed to offer a low-complexity alternative suitable for real-time applications.

In addition to CNNs, this study investigates the Vision Transformer (ViT) model under two training strategies: developing models from scratch and refining pre-trained ones. The inclusion of ViT introduces a transformer-based learning paradigm to the plant disease classification task. A comparative evaluation of all models, focusing on accuracy, generalization ability, and computational efficiency, is presented in the subsequent sections.

Dataset

The experiments in this study are based on the PlantVillage Tomato Leaf Dataset, a publicly available benchmark curated for plant disease classification tasks. The dataset comprises 14,531 high-resolution RGB images of tomato leaves, divided into ten classes: one Healthy class and nine disease categories, including Late Blight, Tomato Mosaic Virus, Septoria Leaf Spot, Leaf Mold, Early Blight, Spider Mites, Target Spot, Bacterial Spot, and Tomato Yellow Leaf Curl Virus.

Each image depicts a single leaf photographed under controlled lighting, usually against a uniform background, in accordance with the PlantVillage data collection protocol. This uniformity supports effective model training while still presenting challenges such as subtle visual differences between disease types.

The dataset includes class-imbalanced distributions, which reflect real-world conditions and require models to generalize effectively across both common and less frequent disease types. During preprocessing, all images were scaled to 224×224 pixels to maintain uniformity across models. Given its clean labelling and high quality, the dataset is widely used for benchmarking classification algorithms in plant pathology.

Transfer Learning-Based Models

In this study, transfer learning is utilized to improve the performance of DL models for classifying tomato leaf diseases. By using deep CNNs pre-trained on large-scale datasets such as ImageNet, transfer learning substantially decreases both training time and reliance on extensive labelled data (Arnob *et al*., 2025). Models such as ResNet50, Inception V3, and VGG16, are fine-tuned on the tomato leaf dataset to leverage their hierarchical feature extraction capabilities. These architectures, originally trained to classify over a thousand general object categories, are repurposed here for the more domain-specific task of plant disease detection.

Transfer learning offers several advantages in this context:

- Utilization of pre-learned weights from large datasets
- Improved generalization even with limited labelled agricultural data
- Faster convergence during training
- The architecture-specific adaptations for this study are described below

VGG16

VGG16 (Simonyan and Zisserman, 2015) is a widely recognized deep convolutional network developed by the VGG group at the University of Oxford. As illustrated in Figure 1, the model comprises 13 convolutional layers and 3 fully connected layers, making up a total of 16 weight layers. Each convolutional layer uses a fixed 3×3 kernel with stride 1 and padding 1, followed by max-pooling layers with a 2×2 window and stride 2 for spatial down sampling. This uniform architecture simplifies implementation while enabling deep hierarchical feature learning.

In our experiments, the top (classification) layers of VGG16 were replaced to suit the 10-class classification task, and the convolutional base was fine-tuned on the tomato leaf dataset.

Inception V3

Inception V3 (Szegedy *et al*., 2016) is an advanced CNN architecture from the GoogleNet family, optimized for both accuracy and computational efficiency. It incorporates Inception modules that apply multiple convolutional filters—1×1, 3×3, and 5×5 in parallel, followed by concatenation. This architecture enables the model to simultaneously learn features at multiple scales, improving its capacity to capture both fine-grained and coarse patterns in images.

The network described in Figure 2 also employs factorized convolutions (e.g., 3×3 into two 1×3 + 3×1) and batch normalization, which collectively reduce the number of parameters and stabilize training. With over 48 layers, Inception V3 is particularly effective for complex image classification tasks.

In this study, the final layers of Inception V3 were replaced with a task-specific head, while the rest of the network was fine-tuned using the tomato disease dataset.

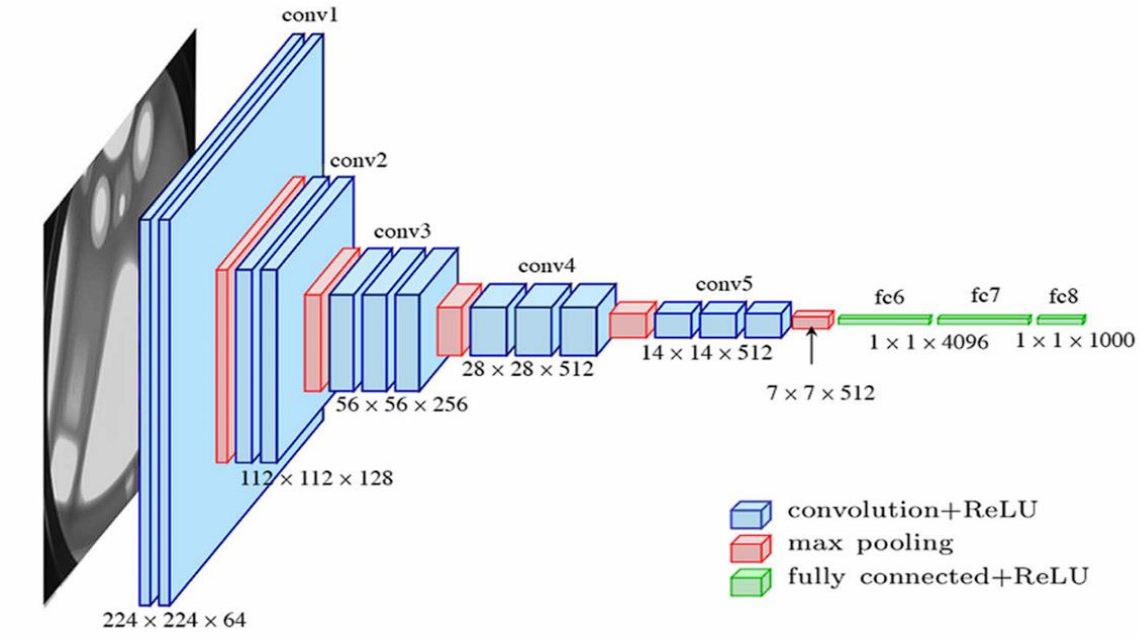


Fig. 1: VGG16 Architecture Overview

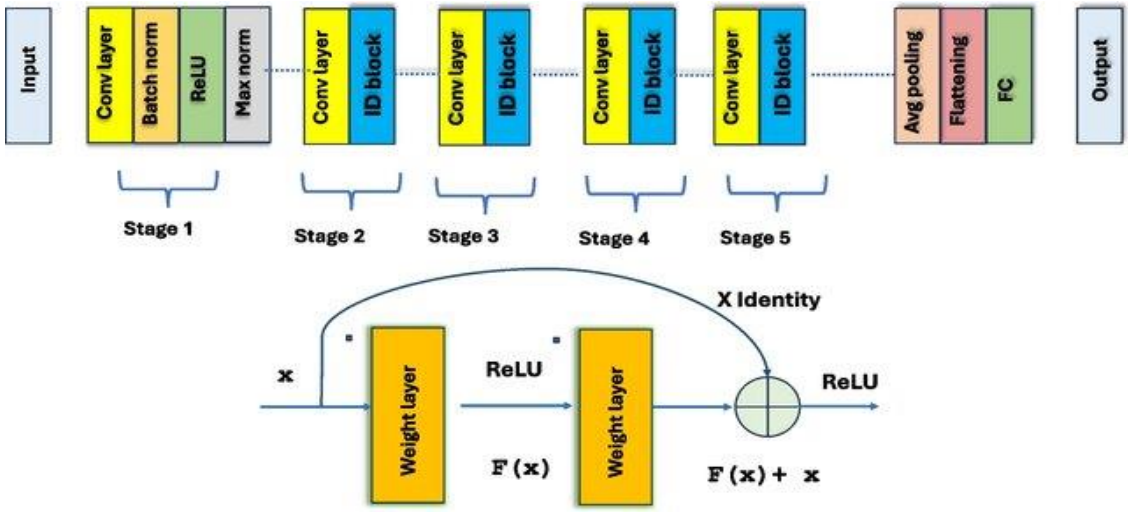


Fig. 2: Illustration of Inception V3 Architecture

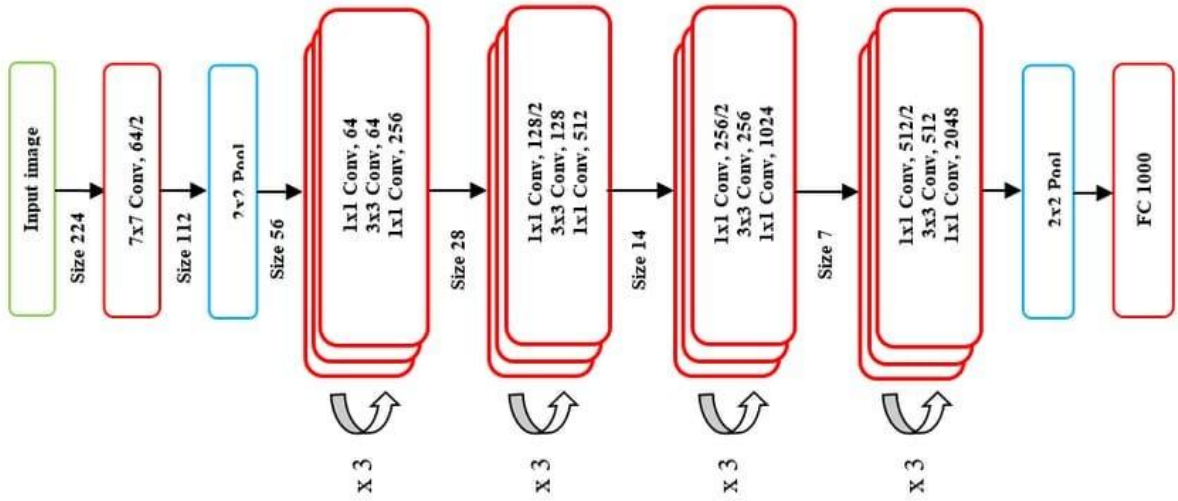


Fig. 3: ResNet50 Architecture

ResNet50

ResNet50 (He *et al*., 2016) is a robust deep CNN architecture that incorporates residual learning, a technique designed to alleviate the vanishing gradient problem in deep networks. It does so through skip (shortcut) connections, which allow the model to learn identity mappings alongside the main transformation. This makes the network easier to optimize, even at depths of 50 layers or more.

The model consists of convolutional blocks and identity blocks, each featuring batch normalization and ReLU activations as presented in Figure 3. The residual connections help maintain gradient flow and enable efficient training of deeper networks, making ResNet50 highly suitable for our task.

Vision Transformer (ViT)

The ViT model, introduced by Dosovitskiy *et al*. (2021), signifies a noteworthy advancement in the application of transformer architectures to computer vision tasks. Inspired by the success of transformers in Natural Language Processing (NLP), ViT adapts the same principles for image classification, fundamentally differing from traditional CNNs (Simonyan and Zisserman, 2015) that count on localized convolutional operations to hierarchically extract spatial features. ViT departs from this paradigm by treating an image as a sequence of fixed-size patches, analogous to word tokens in text, and processes them using a standard transformer encoder architecture.

As illustrated in Figure 4, the image classification pipeline in ViT begins by dividing an input image having $H \times W \times C$ dimensions into non-overlapping patches of size $P \times P$. Each of these image patches is flattened into a vector in one dimension and subsequently projected into a D-dimensional embedding space using a trainable linear projection. These vectors, called patch embeddings, are used as the input sequence for the transformer encoder.

To retain spatial information which is otherwise lost in the flattening process ViT incorporates positional encodings into the patch embeddings. These encodings enable the model to learn the spatial relationships among patches, thereby preserving the global structure of the image. Additionally, a learnable classification token ($[CLS]$) is prepended to the sequence. This token is designed to aggregate contextual information from all patches during transformer processing and is ultimately used for generating the final classification prediction.

These embeddings are then input to a standard Transformer encoder for further processing, which involves multiple layers, each comprising the following key mechanisms:

- Multi-Head Self-Attention (MSA): Enables the model to capture both local and global dependencies by computing attention scores across all patch pairs
- Feedforward Neural Network (FFN): Applies non-linear transformations to the output of the MSA module, enhancing representational power
- Residual connections and layer normalization are employed throughout the encoder to facilitate stable training and improve convergence

After processing by the Transformer encoder, the output corresponding to the ($[CLS]$) token is fed into a fully connected classification head**.** This final layer computes class probabilities via a softmax activation, enabling the model to perform image classification.

Development of a Lightweight CNN

While transfer learning-based models offer robust performance, their high computational complexity and large parameter count make them resource-intensive. To address this limitation, a lightweight CNN architecture is designed specifically for TLDC. Inspired by existing shallow CNN architectures, this model consists of four convolutional layers, each followed by max-pooling layers to reduce feature dimensionality while preserving essential information.

The lightweight CNN architecture employs batch normalization to standardize inputs and accelerate convergence. Dropout is used to reduce overfitting by randomly deactivating neurons, while the final layer uses softmax for multi-class classification.

Architecture of the Customized CNN

The proposed lightweight CNN architecture, detailed in Table 2, includes convolutional layers for feature extraction, max-pooling layers for dimensionality reduction, and a fully connected layer for classification.

The proposed methodology provides a comprehensive framework for TLDC by leveraging pre-trained DL models through transfer learning and introducing a computationally efficient customized CNN architecture.

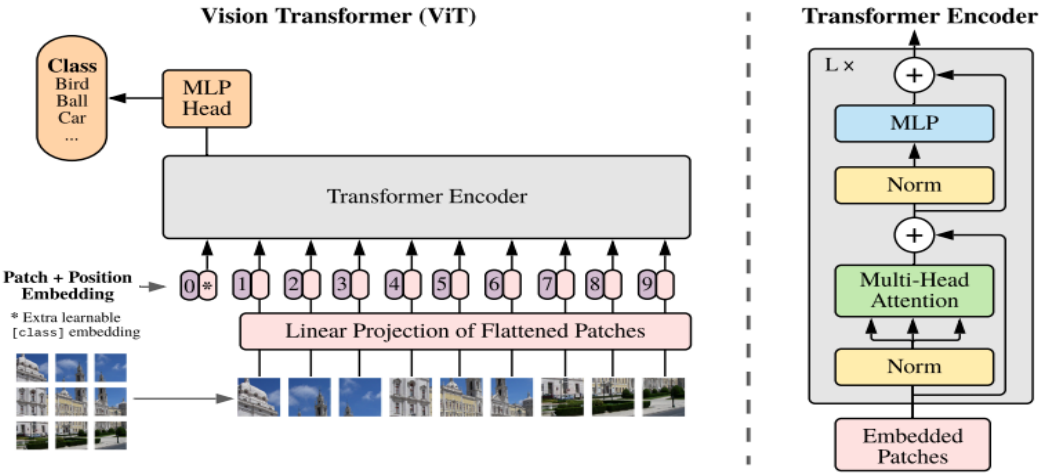


Fig. 4: Vision Transformer architecture

Table 2: Architecture of the Customized CNN

Layer Type	Kernel Size	No. of Kernels	Output Shape	Parameters
Input Layer	-	-	(224,224,3)	0
Conv1	3x3	16	(224,224,16)	448
MaxPool1	2x2	-	(112,112,16)	0
Conv2	3x3	32	(112,112,32)	4,640
MaxPool2	2x2	-	(56,56,32)	0
Conv3	3x3	64	(56,56,64)	18,496
MaxPool3	2x2	-	(28,28,64)	0
Conv4	3x3	128	(28,28,128)	73,856
MaxPool4	2x2	-	(14,14,128)	0
Flatten	-	-	(1,1,25088)	0
Dense-1	-	-	(1,1,512)	12,845,568
Batch Normalization	-	-	(1,1,512)	2,048
Output Layer	-	-	(1,1,10)	5,130
Total Parameters	-	-	-	12,950,186
Trainable Parameters	-	-	-	12,949,162
Non-Trainable Parameters	-	-	-	1,024

Experimental Setup

This section outlines the experimental framework adopted for training and evaluating DL models in the task of TLDC. All experiments were conducted on the Kaggle cloud platform using the PlantVillage Tomato Leaf dataset**,** which comprises a total of 14,531 labeled images**,** categorized into classes including nine disease categories and one healthy.

The dataset was partitioned into training and testing sets using an 80:20 split**,** resulting in 11,624 images for training and 2,907 for testing**.** To further improve model generalization and mitigate overfitting, 10% of the training set (approximately 1,162 images) was reserved as a validation set.

For computational efficiency, an NVIDIA Tesla P100 GPU was utilized to train and fine-tune all DL models. The study employed transfer learning with three pre-trained convolutional architectures Inception V3, VGG16**,** and ResNet50 as well as a custom lightweight CNN developed to suit resource-constrained environments.

In addition to CNN-based models, the Vision Transformer (ViT) architecture was evaluated using two training strategies: training from scratch using only the PlantVillage dataset, and fine-tuning a ViT model pre-trained on ImageNet-21k. All models were trained under a consistent set of hyperparameters and preprocessing steps, summarized in Table 2. Standard metrics, including accuracy, precision, recall, and F1-score, were used to evaluate the models for a reliable comparison.

Model Training and Fine-Tuning

The VGG16 model was adapted using a transfer learning approach, retaining only the final layer as trainable while initializing the remaining layers with weights from the ImageNet-pretrained model. A 10-node dense layer with SoftMax activation was included for the 10-class classification task.

For the Inception V3 model, the original top layer was replaced with two fully connected dense layers, a dropout layer with a rate of 0.2, and a final dense output layer. In the case of ResNet50, a global average pooling layer was added, followed by two dense layers containing 1024 and 10 neurons, respectively.

To assess the effectiveness of the Vision Transformer (ViT) architecture for TLDC, two training strategies were employed. In the first, the ViT model was trained from scratch on the PlantVillage dataset. In the second, a pre-trained ViT model was fine-tuned on the same dataset. Both strategies used the same hyperparameters outlined in Table 3.

Additionally, the customized lightweight CNN model, whose architecture is presented in Table 2, was trained using the same hyperparameters of Table 4 as the pre-trained models.

Table 3: Hyperparameters and Model Specifications for Vision Transformer

Parameter	Value
Patch Size	16 × 16
Optimizer	Adam
Learning Rate	3×10^{-3}
Loss Function	Cross Entropy
Number of Epochs	15

Table 4: Hyperparameters and Model Specifications for lightweight CNN Model

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.001
Loss Function	Categorical Cross-Entropy
Dataset	Tomato Leaf Dataset
No. of Epochs	15
Batch Size	32

Table 5 presents a comparative summary of the five DL Architectures employed in this study. All models were consistent to an input resolution of 224×224 pixels to ensure fair evaluation. As observed, VGG16 possesses the highest number of parameters (138.4 M) owing to its deep stack of convolutional and fully connected layers, while the proposed Lightweight CNN is the most efficient, with only 12.95 M parameters and a shallow architecture comprising just four layers. Inception V3 and ResNet50 maintain a balance between depth and parameter efficiency, with 48 and 50 layers, respectively. The Vision Transformer (ViT-Base) model introduces a transformer-based approach with 12 encoder layers and 86.6M parameters.

Table 5: Comparison of Model Architectures

Model	Input Size	Parameters	No. of Layers
[M1] VGG16	224x224	138.4M	16
[M2] Inception V3	224x224	23.9M	48
[M3] ResNet50	224x224	25.6M	50
[M4] Lightweight CNN	224x224	12.95M	4
[M5] Vision Transformer (ViT-Base)	224x224	86.6 M	12 Transformer Encoder Layers

Results and Discussion

The experimental results obtained from the four DL models are presented. Model performance was assessed using loss metrics, accuracy, F1-score, precision, and recall, with graphical representations illustrating the effectiveness of each model in classifying tomato leaf diseases.

Training and Validation Accuracy

Figure 5 illustrates the validation and training accuracy of all four models. The lightweight CNN, VGG16, and Inception V3 models achieve over 95% training accuracy and more than 85% validation accuracy. The VGG16 model demonstrates slightly higher validation accuracy compared to the other models, making it a strong candidate for TLDC. In contrast, the ResNet50 model exhibits relatively lower performance due to its deeper architecture, which requires extensive training to achieve optimal accuracy.

Fine-tuning the ViT model yielded promising results within 15 epochs, achieving classification accuracy comparable to that of CNNs. The corresponding validation and training accuracy, along with loss curves, for this experiment are illustrated in Figure 6.

Confusion Matrix Analysis

The confusion matrices for all four models, presented in Figure 7, provide a detailed breakdown of classification performance across the 10 disease classes. The results indicate that the VGG16 and lightweight CNN models exhibit strong classification performance, correctly classifying most classes with high confidence. The Inception V3 model also performs well but demonstrates slight misclassifications in some classes. ResNet50, on the other hand, exhibits noticeable misclassifications, particularly in classes with visually similar leaf symptoms, which affects its overall performance.

The confusion matrix for the test dataset, obtained using the fine-tuned ViT approach, is presented in Fig. 8.

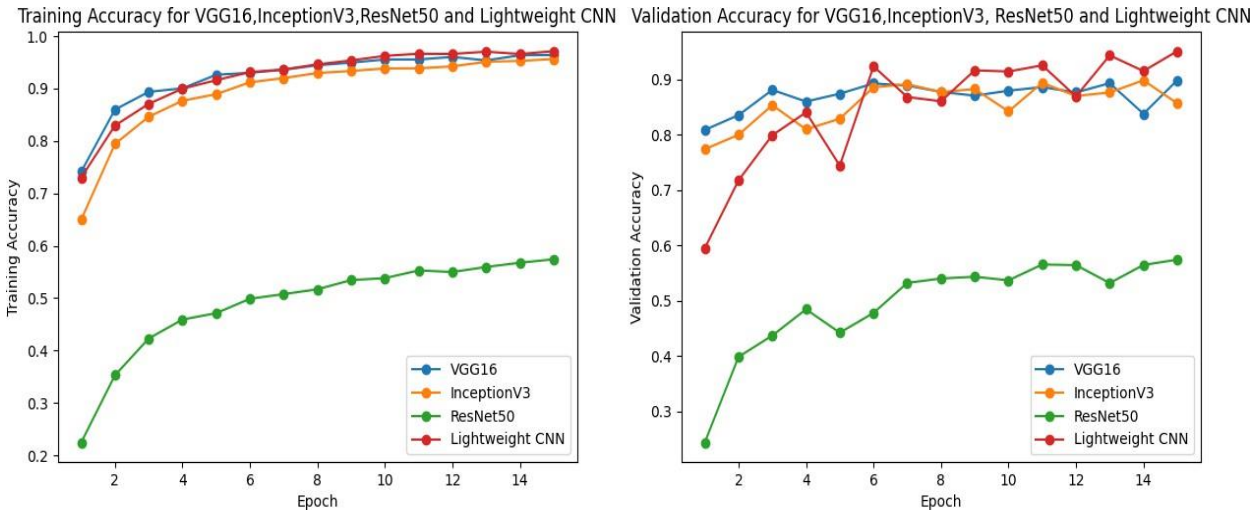


Fig. 5: Training and Validation Accuracy Comparison

DOI: 10.3844/jcssp.2026.47.60

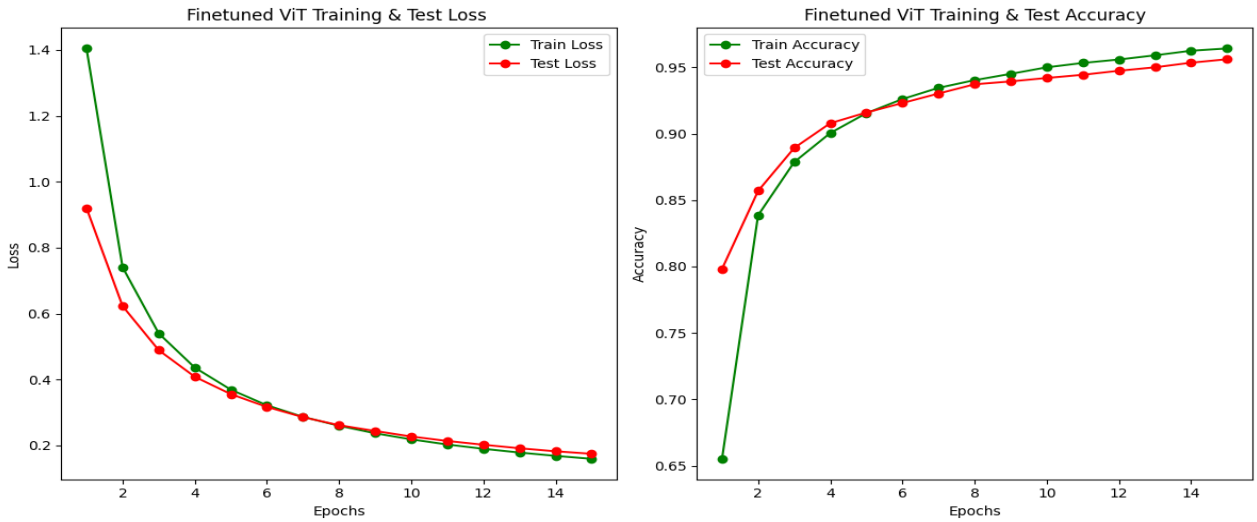
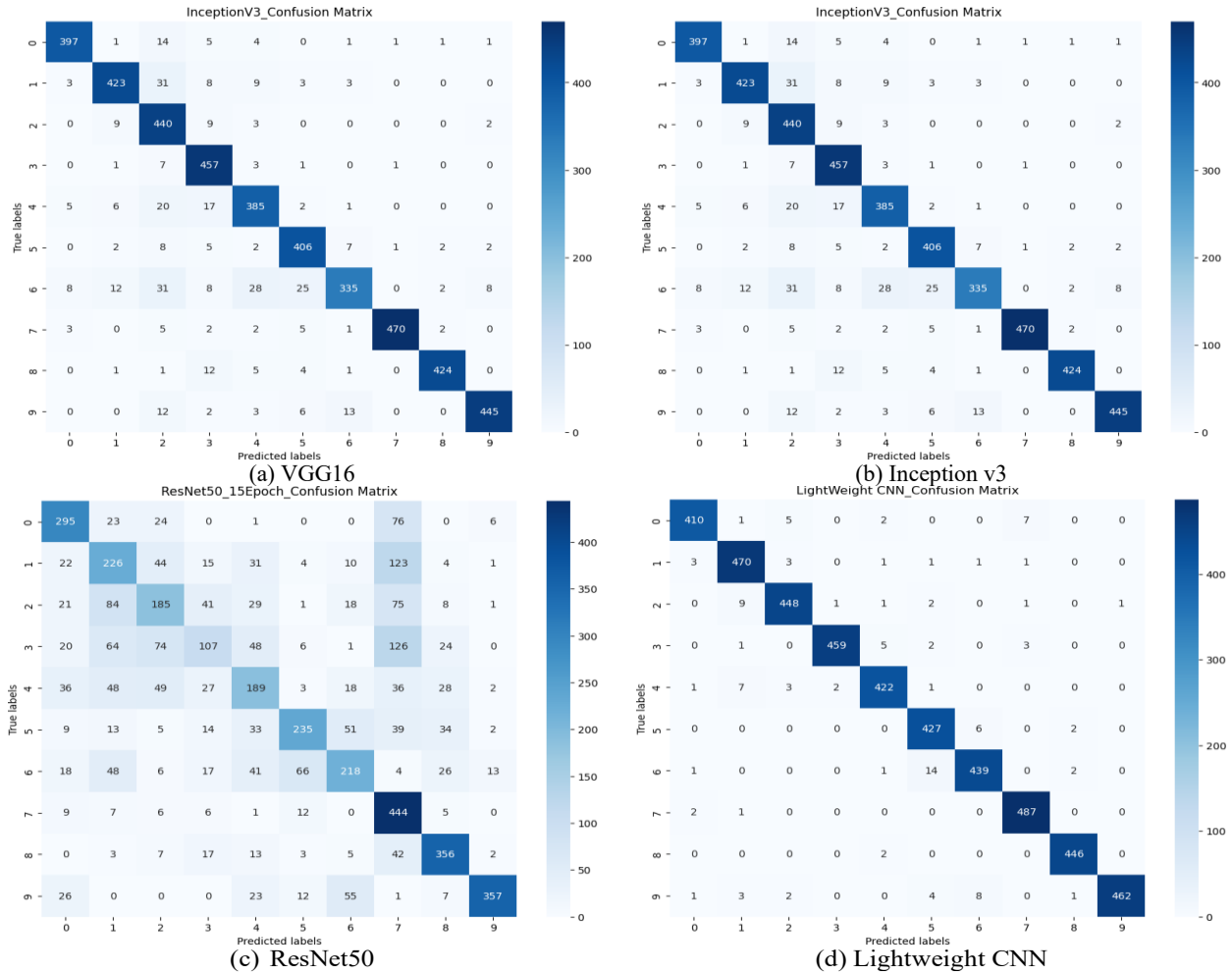


Fig. 6: Training and Validation Accuracy Comparison for Fine Tuned ViT



(a) VGG16 (b) Inception v3

(c) ResNet50 (d) Lightweight CNN

Fig. 7: Confusion Matrices of All Models

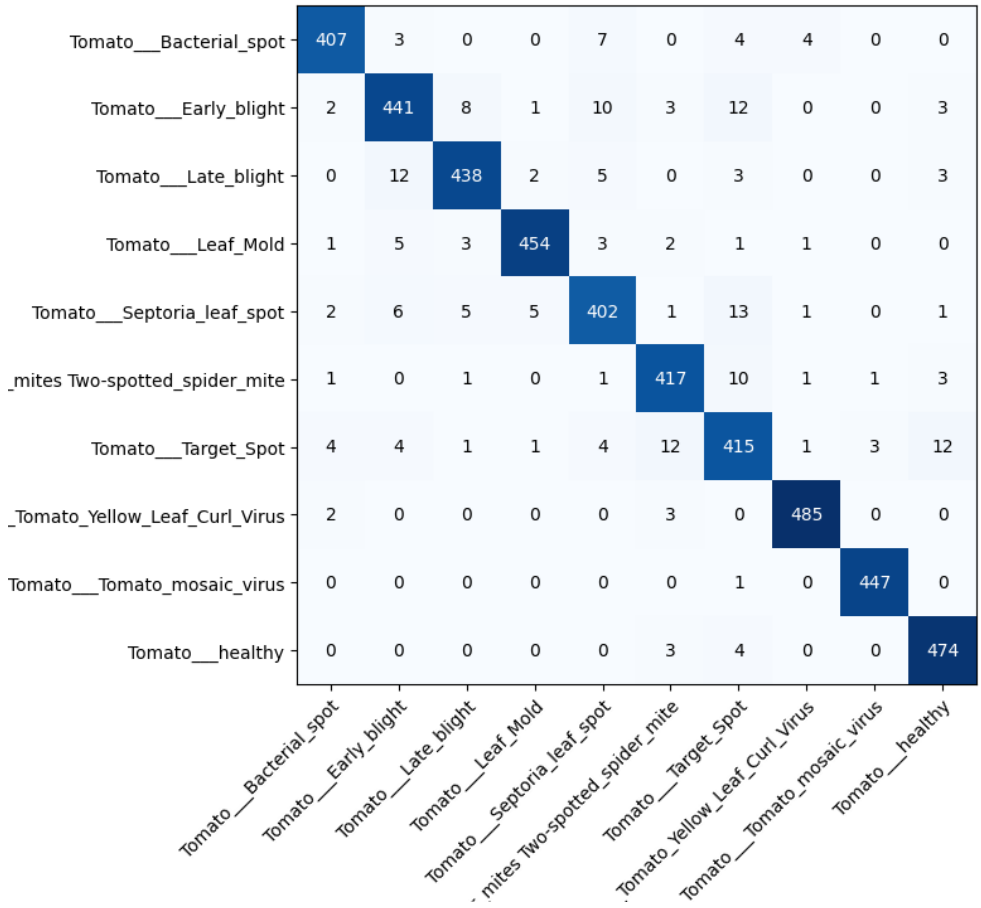


Fig. 8: Confusion Matrix of Finetuned ViT Model

Table 6: Performance Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-Score
[M1] VGG16	93.1%	92.8%	92.5%	92.7%
[M2] Inception V3	88.7%	88.3%	87.9%	88.1%
[M3] ResNet50	84.5%	83.9%	83.2%	83.6%
[M4] Lightweight CNN	92.3%	92.1%	91.8%	92.0%
[M5] Vision Transformer (ViT-Base)	95.6%	95.52%	95.5%	95.4%

Performance Metrics Evaluation

Comprehensive evaluation of the models was carried out by calculating accuracy, precision, recall, and F1-scores. The results, presented in Table 6, reveal that the ViT base model achieved the highest overall classification accuracy and F1-score, outperforming all other architectures considered in the study. It demonstrated superior learning capability, especially in distinguishing between visually similar leaf disease classes, compared to deeper CNN-based models like VGG16, ResNet50, and Inception V3. Additionally, the lightweight CNN architecture showed comparable accuracy while maintaining a minimal number of parameters, making it highly suitable for deployment in computationally constrained settings.

Discussion

This study explored the effectiveness of the ViT architecture for TLDC using the PlantVillage dataset, comparing both a scratch-trained model and a fine-tuned version against established CNN-based architectures. The experimental outcomes clearly indicate that the ViT model, when fine-tuned on domain-specific data, outperforms traditional convolutional approaches in terms of classification accuracy and F1-score.

Training ViT from scratch on the relatively small dataset resulted in severe underfitting and poor performance, reaffirming the model's reliance on large-scale data and high computational resources for effective convergence. Conversely, the fine-tuned ViT

model demonstrated rapid and stable convergence, achieving the highest overall accuracy and generalization ability among all models tested. This highlights the importance of transfer learning in transformer-based architectures, particularly for tasks with limited labeled data.

When compared to deep CNN models such as ResNet50, VGG16, and Inception V3, the fine-tuned ViT model consistently exhibited superior performance across all evaluation metrics. This suggests that ViT's self-attention mechanism is highly effective in capturing global contextual features, which is particularly advantageous for distinguishing subtle inter-class variations in leaf disease patterns.

Moreover, the lightweight CNN architecture, despite its simplicity and significantly lower parameter count, achieved competitive results. Its performance highlights the potential for deploying efficient DL solutions in resource-constrained or edge-computing environments, where computational efficiency is as critical as accuracy.

Overall, the findings of this study not only emphasize the advantages of Vision Transformers in plant disease classification but also underline the trade-offs between model complexity and deployment feasibility. The choice between transformer-based and lightweight CNN architectures can thus be guided by application-specific requirements such as accuracy, latency, and computational constraints.

Conclusion

This research presented a comparative evaluation of the ViT and several CNN architectures for the task of TLDC using the PlantVillage dataset. Among the evaluated models, the fine-tuned ViT base model achieved the highest classification accuracy of 95.53%**,** outperforming established CNN architectures such as VGG16 (91.32%), ResNet50 (92.85%), and Inception V3 (93.67%). These results demonstrate the strong generalization capability of ViT when initialized with pre-trained weights, especially in capturing global context from image patches. In contrast, training ViT from scratch resulted in an accuracy below 10%, underscoring the data-hungry nature of transformer-based models and the necessity of transfer learning for small-scale datasets.

Furthermore, the lightweight CNN model achieved an accuracy of 93.12%**,** offering a competitive alternative with minimal computational requirements. This positions it as a viable candidate for real-time disease detection applications in low-resource environments, such as mobile or embedded agricultural systems.

The findings affirm that transformer-based models, particularly when fine-tuned, hold significant promise for high-accuracy plant disease detection, while lightweight CNNs remain practical for field deployment scenarios.

Future Scope

Despite the promising results achieved in this study, several research avenues remain open for exploration:

- Expansion to Larger and Diverse Datasets: Training on broader, multi-environment datasets will enhance model robustness to real-world variations in lighting, leaf orientation, and disease severity
- Deployment Optimization: Research into model compression techniques (e.g., quantization, pruning, distillation) can help make ViT models feasible for real-time use on edge devices
- Cross-Crop and Multi-Disease Detection: Future work can explore scalable models capable of diagnosing multiple diseases across different crop species using unified frameworks
- Explainable AI Integration: Employing interpretability techniques such as attention heatmaps or saliency maps can improve end-user trust by making model decisions transparent to farmers and agronomists
- IoT-Based Smart Agriculture Systems: Integration of trained models into IoT-based monitoring platforms can enable automated disease surveillance and decision-making support systems for precision farming

By addressing these challenges, future research can significantly enhance the scalability, robustness, and real-world applicability of DL models for plant disease detection.

Acknowledgment

The authors acknowledge the support and facilities provided by the Department of Information Technology, Dharmsinh Desai University. The authors are grateful to the publisher for providing the platform and resources that enabled the dissemination of this research to a wider audience, and they thank the editorial team for their careful review, editorial assistance, and constructive suggestions, which significantly improved the quality and presentation of this paper.

Funding Information

This research received no external funding.

Author's Contributions

Sunil K. Vithlani: Contributed to the conceptualization, methodology design, data collection, data analysis, model development, experimentation, result interpretation, manuscript drafting, and correspondence with the journal.

Vipul K. Dabhi: As the research supervisor, he provided continuous guidance, technical expertise, critical review of the manuscript, and necessary resources throughout the research work.

Ethics

Not applicable, as this study does not involve human participants or animal subjects.

Conflict of Interest

The authors declare that they have no conflicts of interest or competing interests.

Data Availability Statement

The datasets used and analyzed during this study are available from the corresponding author upon reasonable request.

Materials Availability

Not applicable, as no new materials were created or used in this study.

Code Availability

The code used in this research is available from the corresponding author upon reasonable request.

References

Ahmad, A., Saraswat, D., & El Gamal, A. (2023). A Survey on Using Deep Learning Techniques for Plant Disease Diagnosis and Recommendations for Development of Appropriate Tools. *Smart Agricultural Technology*, *3*, 100083. https://doi.org/10.1016/j.atech.2022.100083

Alzahrani, M. S., & Alsaade, F. W. (2023). Transform and Deep Learning Algorithms for the Early Detection and Recognition of Tomato Leaf Disease. *Agronomy*, *13*(5), 1184. https://doi.org/10.3390/agronomy13051184

Amara, J., Bouaziz, B., & Algergawy, A. (2017). A Deep Learning-based Approach for Banana Leaf Diseases Classification. *Datenbanksysteme Für Business, Technologie Und Web (BTW 2017) - Workshopband*, 79–88.

Arnal Barbedo, J. G. (2019). Plant disease identification from individual lesions and spots using deep learning. *Biosystems Engineering*, *180*, 96–107. https://doi.org/10.1016/j.biosystemseng.2019.02.002

Arnob, A. S., Kausik, A. K., Islam, Z., Khan, R., & Bin Rashid, A. (2025). Comparative result analysis of cauliflower disease classification based on deep learning approach VGG16, inception v3, ResNet, and a custom CNN model. *Hybrid Advances*, *10*, 100440. https://doi.org/10.1016/j.hybadv.2025.100440

Bondre, S., & Patil, D. (2024). Crop disease identification segmentation algorithm based on Mask-RCNN. *Agronomy Journal*, *116*(3), 1088–1098. https://doi.org/10.1002/agj2.21387

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv Preprint ArXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Demilie, W. B. (2024). Plant disease detection and classification techniques: a comparative study of the performances. *Journal of Big Data*, *11*(1), 5. https://doi.org/10.1186/s40537-023-00863-9

Du, L., Zhang, R., & Wang, X. (2020). Overview of two-stage object detection algorithms. *Journal of Physics: Conference Series*, *1544*(1), 012033. https://doi.org/10.1088/1742-6596/1544/1/012033

FAOSTAT. (2025). *Food and Agriculture Organization of the United Nations, FAOSTAT*. https://www.fao.org/faostat/en/#home

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587). https://doi.org/10.1109/cvpr.2014.81

Guerrero-Ibañez, A., & Reyes-Muñoz, A. (2023). Monitoring Tomato Leaf Disease through Convolutional Neural Networks. *Electronics*, *12*(1), 229. https://doi.org/10.3390/electronics12010229

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). https://doi.org/10.1109/cvpr.2016.90

Hughes, D. P., & Salathe, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. *ArXiv Preprint ArXiv:1511.08060*. https://doi.org/10.48550/arXiv.1511.08060

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. In *Computers and Electronics in Agriculture* (Vol. 147, pp. 70–90). https://doi.org/10.1016/j.compag.2018.02.016

Kerkech, M., Hafiane, A., & Canals, R. (2020). Vine disease detection in UAV multispectral images using optimized image registration and deep learning segmentation approach. *Computers and Electronics in Agriculture*, *174*, 105446. https://doi.org/10.1016/j.compag.2020.105446

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. *Computer Vision – ECCV 2016*, *9905*, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Li, L., Zhang, S., & Wang, B. (2021). Plant Disease Detection and Classification by Deep Learning—A Review. *IEEE Access*, *9*, 56683–56698. https://doi.org/10.1109/access.2021.3069646

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440. https://doi.org/10.1109/cvpr.2015.7298965

Manjunatha, L., Chowdappa, A., Madhu, G. S., Venkataravanappa, V., Ravikumara, B. M., Ambika, D. S., Keerthi, M. C., Mahadevaiah, C., & Dhanushree, H. K. (2025). Tomato Spotted Wilt Virus. *SpringerLink*, 349–378. https://doi.org/10.1007/978-3-031-81884-4_22

Peyal, H. I., Nahiduzzaman, Md., Pramanik, Md. A. H., Syfullah, Md. K., Shahriar, S. M., Sultana, A., Ahsan, M., Haider, J., Khandakar, A., & Chowdhury, M. E. H. (2023). Plant Disease Classifier: Detection of Dual-Crop Diseases Using Lightweight 2D CNN Architecture. *IEEE Access*, *11*, 110627–110643. https://doi.org/10.1109/access.2023.3320686

Picon, A., Seitz, M., Alvarez-Gila, A., Mohnke, P., Ortiz-Barredo, A., & Echazarra, J. (2019). Crop conditional Convolutional Neural Networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions. *Computers and Electronics in Agriculture*, *167*, 105093. https://doi.org/10.1016/j.compag.2019.105093

Prajapati, H. B., Shah, J. P., & Dabhi, V. K. (2017). Detection and classification of rice plant diseases. *Intelligent Decision Technologies*, *11*(3), 357–373. https://doi.org/10.3233/idt-170301

Karthik, R., Hariharan, M., Anand, S., Mathikshara, P., Johnson, A., & Menaka, R. (2020). Attention embedded residual CNN for disease detection in tomato leaves. *Applied Soft Computing*, *86*, 105933. https://doi.org/10.1016/j.asoc.2019.105933

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. https://doi.org/10.1109/cvpr.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/tpami.2016.2577031

Sakkarvarthi, G., Sathianesan, G. W., Murugan, V. S., Reddy, A. J., Jayagopal, P., & Elsisi, M. (2022). Detection and Classification of Tomato Crop Disease Using Convolutional Neural Network. *Electronics*, *11*(21), 3618. https://doi.org/10.3390/electronics11213618

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ArXiv Preprint ArXiv:1708.08296*. https://doi.org/10.48550/arXiv.1708.08296

Schreinemachers, P., Simmons, E. B., & Wopereis, M. C. S. (2018). Tapping the economic and nutritional power of vegetables. *Global Food Security*, *16*, 36–45. https://doi.org/10.1016/j.gfs.2017.09.005

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv Preprint ArXiv:1409.1556*, *6*. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826. https://doi.org/10.1109/cvpr.2016.308

Trivedi, N. K., Gautam, V., Anand, A., Aljahdali, H. M., Villar, S. G., Anand, D., Goyal, N., & Kadry, S. (2021). Early Detection and Classification of Tomato Leaf Disease Using High-Performance Deep Neural Network. *Sensors*, *21*(23), 7987. https://doi.org/10.3390/s21237987

Wang, G., Sun, Y., & Wang, J. (2017). Automatic Image-Based Plant Disease Severity Estimation Using Deep Learning. *Computational Intelligence and Neuroscience*, *2017*, 1–8. https://doi.org/10.1155/2017/2917536