

Original Research

Received 07 June 2026

Revised 18 July 2026

Accepted 01 August 2026

Natural Resources for Human
Health 2026, Vol. 6 (14s): 1220–1230

KEYWORDS:

Tomato leaf disease classification,
EfficientNet, Vision Transformer,
attention-based fusion, , deep
learning for agriculture, hybrid
CNN-Transformer model

EfficientNet–ViT Hybrid Framework with Attention-Guided Fusion for Tomato Leaf Disease Classification

Pramod P. Gadekar¹, Omkar Patnaik²

¹Department of CSE, School of CSE, Sandip University, Nashik, India ,
Email: mailtoppg18@gmail.com

²Department of CSE, School of CSE, Sandip University, Nashik, India

ABSTRACT: In this paper we provide a hybrid deep learning framework for accurate classification of tomato leaf diseases. The proposed model integrates EfficientNet and Vision Transformer (ViT), where EfficientNet captures fine-grained texture and color variations, and ViT models long-range spatial dependencies. An attention-guided fusion mechanism is introduced to combine these representations by assigning adaptive importance to each branch. The model is evaluated on an eight-class dataset containing 18,160 images, divided into training (70%), validation (15%), and testing (15%) sets. Data augmentation and transfer learning are applied to improve generalization, while Bayesian optimization is used for hyperparameter selection. The proposed approach achieves 98.6% training accuracy, 96.7% validation accuracy, and 96.2% test accuracy, with an F1-score of 0.962 and an AUC of 0.978. Five-fold cross-validation further confirms stable performance with low variance. Comparative results show that the hybrid model outperforms standalone EfficientNet and ViT architectures. These findings demonstrate that combining convolution-based feature extraction with transformer-based global modeling provides a reliable solution for tomato leaf disease classification in practical agricultural applications.

1. INTRODUCTION

Tomato cultivation forms an important part of agricultural production across many regions, and its growth is frequently affected by a wide range of leaf diseases. Conditions such as Early Blight, Late Blight, Leaf Miner damage, nutrient deficiencies, and viral infections can weaken plant health and cause considerable yield loss when not identified at the right stage. Farmers largely depend on visual inspection to detect these problems, but factors like uneven lighting, overlapping leaves, early-stage symptoms, and the subjective nature of manual diagnosis often lead to inconsistent assessments. As cultivation scales up, the limitations of conventional inspection become more evident, creating a clear need for automated and reliable disease recognition systems.

In recent years, deep learning-based image analysis has shown strong potential in crop disease identification. Convolutional Neural Networks (CNNs) have been widely used due to their ability to capture textural patterns, color distortions, and fine structural changes on leaf surfaces. These models have delivered substantial improvements over traditional feature-engineering approaches, yet they carry certain constraints. CNNs are highly effective at extracting local information but often fall short in modeling long-range relationships across the image. This becomes a limitation when disease symptoms are dispersed or when their spatial arrangement carries diagnostic value.

Vision Transformers (ViT) introduced an alternative direction by focusing on global context and interactions between image patches. Their self-attention mechanism enables a more holistic understanding of disease spread, scattered lesions, and shape irregularities. However, ViT models generally require large datasets for stable performance and may overlook subtle texture cues without a complementary feature extractor. Both approaches have individual strengths, but neither fully addresses the

variety of visual patterns present in real-world tomato leaf images when used independently.

Recent work in plant disease detection has moved toward hybrid architectures that combine different learning paradigms to improve robustness. Building on this direction, the present study introduces a hybrid model that integrates EfficientNet and ViT within a unified framework. EfficientNet provides strong local feature extraction with fewer parameters, while ViT contributes a broader understanding of spatial relations. An attention-based fusion layer brings these feature streams together, allowing the system to highlight and combine information from regions that show early or advanced signs of infection.

To support stable training, the system uses transfer learning and a set of augmentation strategies that reproduce realistic variations in field conditions, such as rotation, brightness shifts, and partial occlusion. Class imbalance—common in agricultural datasets—is handled through focal loss to ensure that all disease categories contribute meaningfully to the learning process. The model is trained on a dataset representing eight tomato leaf conditions and evaluated across multiple performance measures.

The goal of this paper is to present a complete and practical hybrid system for tomato leaf disease classification that offers improved accuracy, strong generalization, and clear interpretability. By combining the detailed representation of CNNs with the contextual understanding of Transformers, the proposed model addresses several persistent challenges in digital agriculture and moves toward building reliable automated tools for farmers, agronomists, and precision farming applications.

Unlike existing hybrid models that rely on direct concatenation or fixed fusion strategies, the proposed approach introduces an attention-guided mechanism that dynamically balances local and global features. This improves interpretability and ensures that the contribution of each branch adapts to disease characteristics.

2. RELATED STUDY

Several studies have focused on improving the accuracy and reliability of tomato leaf disease identification, each using different deep learning strategies and image processing techniques.

Kawsar Ahmed et al. developed a Deep Convolutional Neural Network (DCNN) framework that combined image preprocessing with advanced feature extraction to strengthen tomato disease diagnosis. Their method evaluated various filter combinations, including Gaussian and Median filters, along with color model conversions such as HSI and CMYK. Among their experiments, ResNet-50 paired with Gaussian Blur, Gaussian Noise, and RGB-to-CMYK transformation recorded the highest performance, reaching 99.53% accuracy in classifying disease categories [9].

Al Amin Biswas et al. highlighted the need for early and reliable detection to support crop productivity. They designed a lightweight CNN model to classify eleven tomato leaf conditions, covering both diseased and healthy leaves. The approach relied on extensive preprocessing, data augmentation, and transfer learning using VGG-16 and VGG-19. Their final model obtained 95.00% accuracy, and they further deployed it through web and Android interfaces, demonstrating its value for real-time field usage [10].

Kahkashan Perveen et al. proposed a multiscale U-network that focused on segmenting tomato leaf lesions of varying sizes and shapes. To improve the distinction between lesion regions and healthy tissue, the authors introduced multi-scale residual blocks and a Classifier-Bridge (CB) module that connects extracted disease features with segmentation output. The model showed strong resilience across multiple test scenarios, achieving an average accuracy of 99.2%, and remained robust even under reduced brightness and salt-and-pepper noise conditions [11].

Chen et al. refined the AlexNet architecture by integrating Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) using a weighted fusion strategy. Their work focused on eight tomato disease categories and healthy samples, using transfer learning from the PlantVillage dataset. By simplifying and adjusting the fully connected layers, the method managed to improve classification efficiency and reached 98.83% accuracy [12].

Ledbin Vini et al. developed TrioConvTomatoNet, a deep convolution-based model designed to improve feature extraction through a structured series of convolution layers. Using stochastic gradient descent for optimized training and evaluating the model on both existing and real-time datasets, the system demonstrated 99.39% accuracy. The study emphasized its usefulness for early disease identification and its potential for integration into user-friendly applications aimed at supporting farmers [13].

Manjunatha and Jose Mathew employed an enhanced AlexNet known as DbneAlexNet for tomato leaf disease classification. Their workflow began with anisotropic filtering for image refinement, followed by U-Net-based segmentation tuned with Gradient-Golden Search Optimization. A combination of augmentation techniques and Gradient Jaya-Golden optimization during training enabled the system to classify diseases such as bacterial spot and late blight with 92.4% accuracy. However, the authors noted the need for broader dataset validation before deployment in diverse field settings [14].

Shanthi et al. examined the effectiveness of CNN-based architectures, particularly AlexNet and VGG-16, for early-stage detection of tomato leaf diseases. Their experiments showed that CNNs play an important role in improving yield and reducing losses when deployed at an early growth stage. Their model achieved 91.2% accuracy, but the study acknowledged the need for more diverse data and further architectural refinement to improve overall robustness [15].

Neeraj et al. used YOLOv5 and YOLOv7 models enhanced with transfer learning and ensemble strategies to detect tomato leaf diseases. The combined YOLO approach recorded 98.8% accuracy, with high precision and recall values. The main challenges reported were increased computational load and limited dataset diversity, both of which were partially addressed through transfer learning [16].

Santar Pal Singh et al. adopted a modified Mask R-CNN model to simultaneously detect and segment tomato leaf diseases. Their method achieved 98% accuracy, showing good efficiency in identifying lesion regions. However, the authors pointed out that accurate annotations and misidentification of complex lesion patterns remain limitations that need attention for large-scale use [17].

Table 1 summarizes the major contributions from existing literature, outlining dataset size, number of classes, model type, and reported accuracy. Together, these studies reveal the growing trend toward deep learning-based agricultural diagnostics. Building upon the gaps identified—such as limited generalization, high computational cost, and challenges in capturing both local and global patterns—the present work employs a hybrid EfficientNet–ViT architecture with an attention-based fusion mechanism to enhance tomato leaf disease classification.

Table 1. Comparative Analysis

Method or Model Used	Dataset Size	No. of Classes	Accuracy (%)
DbneAlexNet with anisotropic filtering and U-Net segmentation [14]	18,161	10	92.40
TrioConvTomatoNet (deep CNN with staged convolution layers)[13]	18,160	10	99.39
DCNN with Gaussian/Median filtering and CMYK conversion (ResNet-50 best)[9]	6,794	3	99.53
Lightweight custom CNN with TL (VGG-16/19)[10]	32,535	11	95.00
Multiscale U-Net with CB module[11]	18,096	10	99.20
Enhanced AlexNet with HOG–LBP feature fusion[12]	13,038	9	98.83
CNN (AlexNet, VGG-16) [15]	2,782	6	91.20
Combined YOLOv5 + YOLOv7 (ensemble)[16]	18,162	10	98.80
Modified Mask R-CNN[17]	1,610	7	98.00

Reedha et al. [18] explored the use of a convolution-free Vision Transformer (ViT) for tomato leaf disease classification. Their approach relied purely on self-attention, treating each image as a sequence of fixed patches processed by a standard transformer encoder. Even though the dataset was relatively small, the authors achieved high classification accuracy. They attributed this performance to strong data augmentation, the advantage of transfer learning, and the limited number of target classes, which collectively reduced the model’s risk of overfitting.

Thai et al. [19] applied a ViT-based architecture for cassava leaf disease recognition and reported accuracy gains of at least 1% over several well-known CNN models. Their model also reached a 90% F1-score, confirming its consistency across classes. However, the system contained 85.79 million parameters, making it unsuitable for edge or IoT devices. To overcome this limitation, they quantised the model, achieving a threefold reduction in size before deploying it on the Raspberry Pi 4 platform.

Several other studies have proposed variations of ViT to improve performance, efficiency, and adaptability for agricultural imaging applications. In [20], the authors introduced TLMViT, a hybrid design that combines transfer learning with ViT. Pre-trained CNN models such as VGG19 and ResNet50 were used as initial feature extractors, after which a ViT module was applied in deeper layers. This model performed strongly on both the PlantVillage dataset (98.81% accuracy) and a wheat disease dataset (99.86% accuracy), showing the benefits of combining convolutional features with transformer-based global reasoning.

Recent work has also demonstrated the effectiveness of the Swin Transformer for plant disease classification. Sun et al. [21] evaluated a Swin Transformer-based method using the PlantVillage and Tomato-Village datasets. Their system reached 99.7% accuracy, and the addition of a CNN-attention module further improved performance, confirming the strength of hierarchical attention for fine-grained symptom recognition.

Building on the trend toward lightweight models for IoT deployment, Thakur et al. [22] introduced PlantXViT, a compact hybrid architecture that integrates CNN layers with ViT blocks. Designed for smart agriculture settings, PlantXViT delivered accuracies of 93.5%, 92.5%, and 98.3% on apple, maize, and rice datasets, respectively. Its favorable trade-off between accuracy and model size makes it suitable for real-time field applications.

MobileViT represents another important direction, blending CNN efficiency with ViT's global attention capabilities. Han et al. [23] proposed an improved MobileViT version that incorporates Squeeze-and-Excitation modules and a global attention mechanism. Their enhanced design improved feature fusion and representation quality, achieving 88.86% accuracy, which is a 4.28% increase over the original MobileViT model.

Tonmoy et al. [24] further advanced this line of work by introducing MobilePlantViT, a lightweight hybrid ViT model designed specifically for plant disease classification. Despite having only 0.69 million parameters, the model outperformed larger variants like MobileViTv1 and MobileViTv2, achieving test accuracies ranging from 80% to over 99% across multiple datasets. The study demonstrated that well-designed hybrid transformer models can remain highly accurate even with extremely low parameter counts, making them suitable for resource-limited agricultural systems.

Hossain et al. [25] performed a comparative study of transformer-based architectures for tomato leaf disease classification and found that MaxViT delivered the strongest performance with an accuracy of 97%. Meanwhile, Emmanuel and Hidayatullah [26] evaluated a series of pre-trained ViT models, including EfficientViT-b, EfficientViT-m, and MobileViT. Their results showed that EfficientViT-b2 offered the best overall accuracy for plant disease classification, provided that computational cost and training time are not restrictive.

Existing studies demonstrate strong performance using CNN, transformer, and hybrid approaches. However, CNN-based models are limited in capturing global dependencies, while transformer-based models often require large datasets and high computational cost. Many hybrid models rely on simple fusion strategies without adaptive weighting, which may reduce effectiveness. These limitations motivate the design of the proposed attention-guided hybrid framework.

3. PROPOSED SYSTEM

The proposed framework is designed to capture both fine-grained local features and global spatial relationships present in tomato leaf disease images. The overall pipeline consists of four stages: (1) dataset preparation and preprocessing, (2) hybrid feature extraction using EfficientNet and Vision Transformer (ViT), (3) attention-guided feature fusion, and (4) classification and optimization.

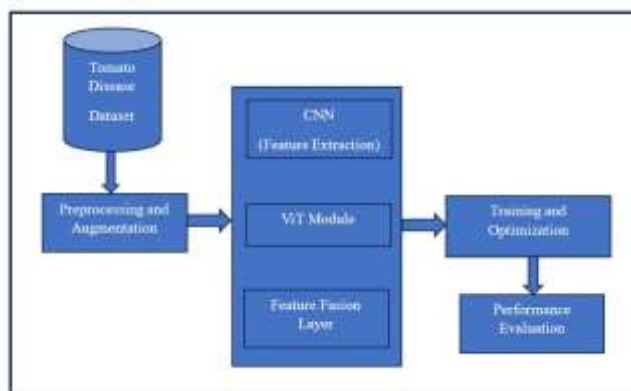


Figure 2. Proposed system architecture

3.1 Dataset Preparation and Preprocessing

The dataset used in this study consists of 18,160 tomato leaf images categorized into eight classes, including healthy leaves and various disease conditions such as fungal infections, pest damage, and viral symptoms. The dataset was compiled from publicly available agricultural image repositories, including PlantVillage and its extended variants.

To ensure reproducibility, the dataset was divided as follows:

- Training set: 70%
- Validation set: 15%
- Test set: 15%

The class distribution is moderately imbalanced, with sample counts ranging between approximately 1,800 and 2,600 images per class. All images were resized to $224 \times 224 \times 3$, and pixel values were normalized to the range $[0, 1]$.

Data Augmentation

To simulate real-world variability and reduce overfitting, the following augmentation techniques were applied:

- Rotation ($\pm 20^\circ$)
- Brightness adjustment
- Random zoom and shear
- Horizontal flipping
- Random erasing (10–20% region masking)

3.2 Hybrid Model Architecture

The proposed model combines EfficientNetB0 and Vision Transformer (ViT) in a dual-branch architecture to capture complementary feature representations

3.2.1 EfficientNet Backbone for Local Feature Extraction

EfficientNetB0 is used as the convolutional backbone for extracting local features such as:

- texture variations
- lesion boundaries
- color distortions

Training strategy:

- Initial layers ($\approx 70\%$) are frozen during early epochs
- Gradual fine-tuning is applied in later stages

3.2.2 Vision Transformer Branch

The ViT branch processes the input image as a sequence of fixed-size patches (16×16). Each patch is embedded and passed through transformer encoder layers to model global relationships.

The ViT captures:

- long-range dependencies
- spatial distribution of disease patterns
- contextual interactions between regions

Training strategy:

- Final transformer layers are fine-tuned
- Pretrained ImageNet weights are used for initialization

3.3 Attention-Guided Feature Fusion

Let $F_c \in \mathbb{R}^d$ denote the feature vector from EfficientNet and $F_t \in \mathbb{R}^d$ denote the feature vector from the Vision Transformer.

Both feature vectors are projected into a common space:

$$\hat{F}_c = W_c F_c, \hat{F}_t = W_t F_t$$

Attention weights are computed as:

$$\alpha = \text{softmax}(W_a[\hat{F}_c; \hat{F}_t])$$

The final fused representation is obtained as:

$$F = \alpha \cdot \hat{F}_c + (1 - \alpha) \cdot \hat{F}_t$$

3.4 Classification Layer

The fused feature representation is passed through:

- Dense layer with 512 units (ReLU activation)
- Dropout layer to reduce co-adaptation
- Dense layer with 256 units
- Final **Softmax layer** producing probabilities for 8 disease classes

3.5 Model Training and Optimization

Transfer Learning

Both branches are initialized with ImageNet weights to accelerate convergence.

Hyperparameter Optimization

Bayesian optimization is used with the following search space:

- Learning rate: 10^{-5} to 10^{-3}
- Batch size: {16, 32, 64}
- Dropout: 0.2–0.5

Loss Function

Focal loss is used to address class imbalance:

$$FL(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t)$$

Training Setup

- Optimizer: Adam
- Epochs: 20
- Early stopping based on validation loss

4. RESULTS AND DISCUSSION

The performance of the proposed EfficientNet–ViT hybrid model was evaluated on the test set after training for 20 epochs. In addition to training and validation performance, the model was assessed using accuracy, F1-score, and AUC to ensure a comprehensive evaluation. To verify robustness, five-fold cross-validation was also conducted, and comparative experiments were performed against standalone EfficientNet and Vision Transformer models.

4.1 Training and Validation Performance

The model showed stable convergence during training. Accuracy increased from **74%** in the first epoch to **98.6%** at the final epoch, while validation accuracy reached **96.7%**. Both training and validation losses decreased consistently and converged to approximately **0.09**, indicating effective learning without significant overfitting.

Table 2: Accuracy over different epoch

Epoch	Training Accuracy	Validation Accuracy
1	0.74	0.71
5	0.85	0.83
10	0.92	0.91
15	0.96	0.95
20	0.98	0.96

4.2 Test Performance

The final evaluation on the held-out test set demonstrates the generalization capability of the proposed model. The model achieved an accuracy of 96.2%, an F1-score of 0.962, and an AUC of 0.978 as summarized in Table 2. These results confirm that the model maintains high performance on unseen data and does not rely on overfitting to the training distribution.

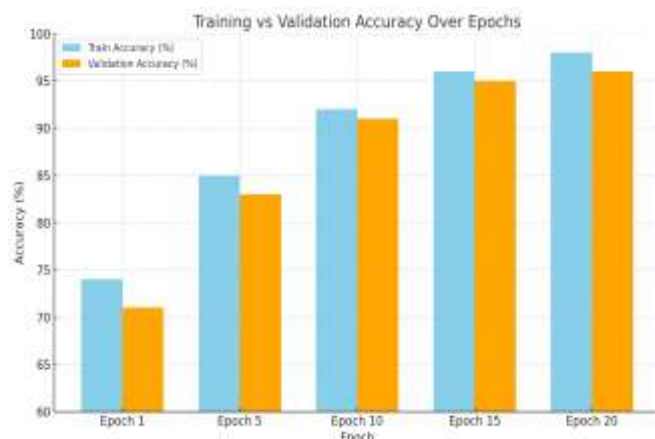


Figure 3: Training Accuracy and validation accuracy

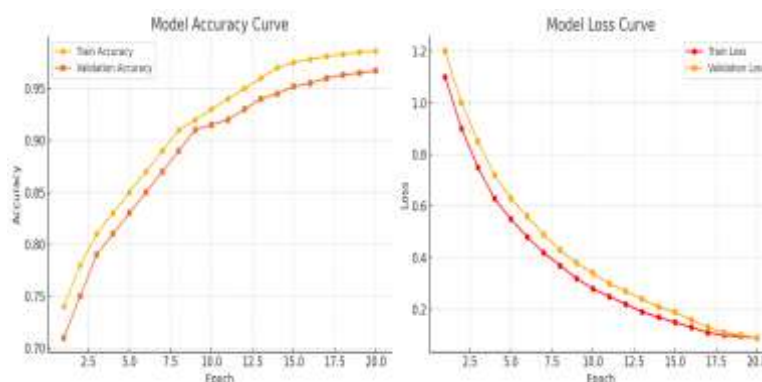


Figure 4: Accuracy and Loss Curve

4.3 Cross-Validation Analysis

To further assess robustness, five-fold cross-validation was performed. The model achieved a mean accuracy of 95.9% with a standard deviation of 0.6, indicating stable performance across different data splits and confirming the reliability of the proposed approach.

4.4 Comparative Evaluation

To evaluate the effectiveness of the hybrid architecture, comparative experiments were conducted using standalone EfficientNetB0 and Vision Transformer models under the same training conditions.

Table 3: Comparative Analysis

Model	Accuracy	F1-score
EfficientNetB0	94.1%	0.941
Vision Transformer	93.5%	0.935
Proposed Hybrid Model	96.2%	0.962

The hybrid model outperforms both individual architectures, demonstrating that combining local feature extraction with global contextual modeling improves classification performance.

4.5 Confusion Matrix Analysis

The confusion matrix indicates strong class-wise performance across all eight categories. Most classes are correctly classified with minimal errors. Misclassifications are primarily observed between visually similar categories, such as nutrient deficiencies. The attention-guided fusion mechanism improves discrimination by focusing on disease-relevant regions, reducing ambiguity between similar patterns.

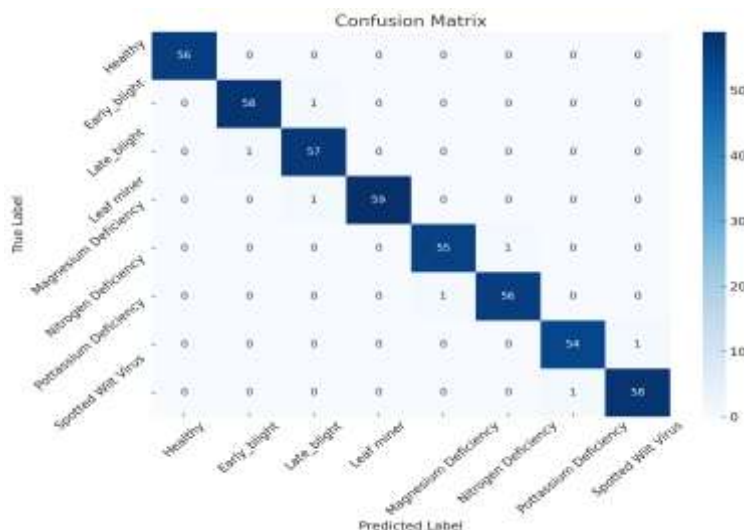


Figure 5: Confusion Matrix

4.6 Model Complexity and Deployment Considerations

The proposed model contains approximately 28 million parameters and achieves an average inference time of 42 ms per image on GPU hardware. Although the hybrid design introduces additional computational cost compared to standalone CNN models, the performance gain justifies its use. The model can be further optimized using pruning or quantization techniques for deployment on edge devices or mobile platforms.

5. CONCLUSIONS

This study presented a hybrid deep learning framework that combines EfficientNet and Vision Transformer modules for tomato leaf disease classification. The system was designed to capture both local texture variations and broader spatial relationships, which are essential for identifying diseases that exhibit scattered lesions, subtle discoloration, or complex symptom patterns. An attention-driven fusion layer was introduced to bring together the strengths of both branches, ensuring that the model concentrates on the most informative regions of each image.

The experimental results demonstrate that the hybrid approach delivers high accuracy and stable performance across all eight disease categories. The validation accuracy reached 96.7%, and the confusion matrix indicates consistently correct classification with only minor errors. The convergence of training and validation curves shows that the model generalizes well and handles class imbalance effectively through focal loss and augmentation strategies.

Overall, the findings highlight the value of integrating CNN-based detail extraction with Transformer-based global understanding for agricultural image analysis. By strengthening the reliability of disease recognition under varied imaging conditions, the proposed system offers practical benefits for precision farming and automated crop monitoring. Future work can focus on extending the dataset with field-captured images, incorporating severity estimation, and adapting the model for deployment on mobile or edge devices to support real-time diagnosis in farming environments.

REFERENCES

- [1] Nishankar, S., Sathiyamohan, S., et al. "ViT-RoT: Vision Transformer-Based Robust Framework for Tomato Leaf Disease Recognition." *AgriEngineering*, vol. 7, no. 6, 2025, p. 185.

- [2] Shehu, H.A., et al. "Early Detection of Tomato Leaf Diseases Using Transformers and Transfer Learning." *European Journal of Agronomy*, vol. 168, 2025, p. 127625.
- [3] Chen, M., et al. "Tomato Leaf Disease Detection Method Based on Improved YOLOv8n." *Scientific Reports*, vol. 15, no. 1, 2025, p. 25837.
- [4] Ouamane, A., et al. "Optimized Vision Transformers for Superior Plant Disease Detection." *IEEE Access*, 2025.
- [5] Swapno, S.M.M.R., et al. "ViT-SENet-Tom: A Novel Hybrid Squeeze–Excitation Network and Vision Transformer Framework for Tomato Fruit Classification." *Neural Computing and Applications*, vol. 37, no. 9, 2025, pp. 6583–6600.
- [6] Vengaiah, C., and Konda, S.R. "A Comparative Study of Convolutional Neural Network Architectures for Enhanced Tomato Leaf Disease Classification Using Refined Statistical Features." *Traitement du Signal*, vol. 41, no. 1, 2024, pp. 201–212.
- [7] Özyurt, F., Sert, E., and Avci, D. "Ensemble Residual Network Features and Cubic-SVM Based Tomato Leaves Disease Classification System." *Traitement du Signal*, vol. 39, no. 1, 2022, pp. 71–77.
- [8] Bayram, H.Y., Bingol, H., and Alatas, B. "Hybrid Deep Model for Automated Detection of Tomato Leaf Diseases." *Traitement du Signal*, vol. 39, no. 5, 2022, pp. 1781–1787.
- [9] Hossain, M.I., Jahan, S., Al Asif, M.R., Samsuddoha, M., and Ahmed, K. "Detecting Tomato Leaf Diseases by Image Processing Through Deep Convolutional Neural Networks." *Smart Agricultural Technology*, vol. 5, 2023, p. 100301.
- [10] Paul, S.G., Biswas, A.A., Saha, A., Zulfiker, M.S., Ritu, N.A., Zahan, I., Rahman, M., and Islam, M.A. "A Real-Time Application-Based Convolutional Neural Network Approach for Tomato Leaf Disease Classification." *Array*, vol. 19, 2023, p. 100313.
- [11] Perveen, K., Debnath, S., Pandey, B., Chand, S.P., Bukhari, N.A., Bhowmick, P., Alshaikh, N.A., Arzoo, S., and Batool, S. "Deep Learning-Based Multiscale CNN-Based U-Net Model for Leaf Disease Diagnosis and Lesion Segmentation in Tomato." *Physiological and Molecular Plant Pathology*, vol. 128, 2023, p. 102148.
- [12] Qiu, J., Lu, X., Wang, X., Chen, C., Chen, Y., and Yang, Y. "Research on Image Recognition of Tomato Leaf Diseases Based on Improved AlexNet Model." *Heliyon*, vol. 10, no. 13, 2024.
- [13] Vini, S.L., and Rathika, P. "TrioConvTomatoNet: A Robust CNN Architecture for Fast and Accurate Tomato Leaf Disease Classification for Real-Time Application." *Scientia Horticulturae*, vol. 330, 2024, p. 113079.
- [14] Badiger, M., and Mathew, J.A. "Tomato Plant Leaf Disease Segmentation and Multiclass Disease Detection Using Hybrid Optimization Enabled Deep Learning." *Journal of Biotechnology*, vol. 374, 2023, pp. 101–113.
- [15] DL, S., K, V., N, A., and Vashistha, S. "Tomato Leaf Disease Detection Using CNN." 2024.
- [16] Kumar, N.S., Sony, J., Premkumar, A., Meenakshi, R., and Nair, J.J. "Transfer Learning-Based Object Detection Models for Improved Diagnosis of Tomato Leaf Disease." *Procedia Computer Science*, vol. 235, 2024, pp. 3025–3034.
- [17] Kaur, P., Harnal, S., Gautam, V., Singh, M.P., and Singh, S.P. "An Approach for Characterization of Infected Area in Tomato Leaf Disease Based on Deep Learning and Object Detection Technique." *Engineering Applications of Artificial Intelligence*, vol. 115, 2022, p. 105210.
- [18] Reedha, R., Dericquebourg, E., Canals, R., and Hafiane, A. "Transformer Neural Network for Weed and Crop Classification of High-Resolution UAV Images." *Remote Sensing*, vol. 14, 2022, p. 592.
- [19] Thai, H.T., Tran-Van, N.Y., and Le, K.H. "Artificial Cognition for Early Leaf Disease Detection Using Vision Transformers." *Proceedings of the International Conference on Advanced Technologies for Communications (ATC)*, Ho Chi Minh City, Vietnam, 14–16 October 2021, pp. 33–38.
- [20] Tabbakh, A., and Barpanda, S.S. "A Deep Features Extraction Model Based on the Transfer Learning Model and Vision Transformer (TLMViT) for Plant Disease Classification." *IEEE Access*, vol. 11, 2023, pp. 45377–45392.

- [21] Sun, Y., Ning, L., Zhao, B., and Yan, J. “Tomato Leaf Disease Classification by Combining EfficientNetV2 and a Swin Transformer.” *Applied Sciences*, vol. 14, 2024, p. 7472.
- [22] Thakur, P.S., Khanna, P., Sheorey, T., and Ojha, A. “Explainable Vision Transformer Enabled Convolutional Neural Network for Plant Disease Identification: PlantXViT.” *arXiv*, 2022, arXiv:2207.07919.
- [23] Han, Z., and Sun, J. “Tomato Leaf Diseases Recognition Model Based on Improved MobileViT.” *Proceedings of the 2024 IEEE 4th International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, Chongqing, China, 6–8 December 2024, pp. 1205–1209.
- [24] Tonmoy, M.R., Hossain, M.M., Dey, N., and Mridha, M. “MobilePlantViT: A Mobile-Friendly Hybrid ViT for Generalized Plant Disease Image Classification.” *arXiv*, 2025, arXiv:2503.16628.
- [25] Hossain, S., Reza, M.T., Chakrabarty, A., and Jung, Y.J. “Aggregating Different Scales of Attention on Feature Variants for Tomato Leaf Disease Diagnosis from Image Data: A Transformer-Driven Study.” *Sensors*, vol. 23, 2023, p. 3751.
- [26] Hamdi, E.B.; Hidayaturrehmana. Ensemble of pre-trained vision transformer models in plant disease classification, an efficient approach. *Procedia Comput. Sci.* 2024, 245, 565–573