Explainable Deep Neural Network Based Multiclass Classification of Tomato Leaf Disease

Rupesh Kumar[1], Dr. Ranu Pandey[2]
[1, 2]Dept of Computer Science
[1, 2] Sri RawatpuraSarkar University, Raipur, India

Abstract- *Tomato leaf diseases pose a persistent challenge to sustainable crop production, requiring accurate and computationally efficient diagnostic solutions for real-time field applications. This study proposes a hybrid deep learning framework that combines a lightweight convolutional neural network (CNN) with lightweight transformer architecture for robust classification of tomato leaf diseases. The CNN component is employed to extract fine-grained local spatial features, while the transformer module captures global contextual dependencies, enabling improved discrimination between visually similar disease categories.*

The model was trained and evaluated on a multi-class tomato leaf dataset comprising healthy and diseased samples, including bacterial spot, early blight, late blight, leaf mold, and septoria leaf spot. Data augmentation and transfer learning strategies were applied to enhance generalization and mitigate overfitting. The proposed hybrid model achieved an overall classification accuracy of 99.08%, with a precision of 98.93%, recall of 98.95%, and F1-score of 98.94%. Comparative analysis indicates superior performance over standalone lightweight CNN and transformer models, while maintaining reduced computational complexity suitable for resource-constrained environments.

To improve model interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) was utilized to visualize disease-relevant regions, confirming that the model focuses on meaningful pathological features. The results demonstrate that the proposed approach provides a reliable, efficient, and interpretable solution for automated plant disease detection, supporting its applicability in precision agriculture and smart farming systems.

Keywords: Tomato leaf disease detection, Hybrid deep learning, Lightweight convolutional neural network, Lightweight Vision transformer, Precision agriculture, Explainable artificial intelligence, Grad-CAM, Image classification.

I. INTRODUCTION

Agriculture continues to play a crucial role in sustaining global food systems, especially in developing regions where crop productivity directly influences economic stability. Among widely cultivated crops, tomato holds significant importance due to its high nutritional and commercial value. However, tomato plants are highly susceptible to a variety of leaf diseases, including bacterial spot, early blight, late blight, leaf mold, and septoria leaf spot. If these diseases are not identified at an early stage, they can lead to considerable yield losses. Traditionally, disease identification relies on visual inspection by agricultural experts, which is often time-consuming, subjective, and difficult to scale in real-world farming conditions.

With the rapid progress of artificial intelligence, deep learning has emerged as a powerful tool for solving complex image classification problems [1]. The availability of large, annotated datasets such as PlantVillage has further enabled the development of automated plant disease detection systems [2]. Early deep learning models, including AlexNet [3] and VGGNet [4], demonstrated the potential of convolutional neural networks (CNNs) in extracting meaningful features from images. These approaches were soon adopted in agriculture, where CNN-based models achieved promising results in plant disease classification tasks [5].

Over time, more advanced CNN architectures such as ResNet [6] and Xception [7] were introduced, addressing limitations like vanishing gradients and improving feature representation. These improvements led to better performance in identifying plant diseases, including those affecting tomato leaves [8]. At the same time, the concept of attention mechanisms brought a new direction to deep learning research. Initially proposed for sequence modeling [9], transformers were later adapted for computer vision tasks, resulting in models such as Vision Transformer (ViT) [19] and Swin Transformer [20]. These architectures are particularly effective in capturing global relationships within images, which can be beneficial for distinguishing visually similar disease patterns.

Despite these advancements, one of the key challenges in deploying deep learning models in agriculture is their lack of interpretability. Farmers and practitioners often require an understanding of how a model arrives at a particular decision. Techniques such as Grad-CAM provide visual explanations by highlighting the regions in an image that contribute most to the prediction [10]. Other explainable AI approaches, including SHAP [14] and broader interpretability frameworks [23], have also been explored to improve transparency and trust in AI systems.

Another important consideration is computational efficiency. In practical scenarios, especially in rural or resource-limited settings, models need to be lightweight and efficient enough to run on mobile or edge devices. Architectures such as MobileNet [11] and MobileNetV2 [12] were specifically designed for such environments, while EfficientNet [16] and EfficientNetV2 [21] further improved the balance between accuracy and computational cost. Additionally, datasets like PlantDoc have contributed to making models more robust by incorporating real-world variations such as complex backgrounds and lighting conditions [22].

Even with these developments, standalone CNNs and transformer models have their own limitations. CNNs are well-suited for capturing local features but may struggle to model long-range dependencies, whereas transformers excel at capturing global context but often require large datasets and higher computational resources. To overcome these limitations, recent studies have explored hybrid approaches that combine CNNs with transformer architectures. Models such as PlantXViT [24] and TomFormer [28] demonstrate how integrating these two paradigms can lead to improved performance in plant disease detection. Other recent works have also reported performance gains using hybrid and fusion-based techniques [29–31].

More recently, researchers have focused on improving both accuracy and robustness through ensemble and hybrid learning strategies [33], while lightweight CNN–transformer combinations have shown promising results for efficient deployment [34]. In addition, the integration of explainability into such models has gained attention, as seen in approaches like XSE-TomatoNet, which combines high-performance classification with interpretability for tomato leaf diseases [35].

Building on these insights, this study proposes a hybrid lightweight deep learning framework that combines a convolutional neural network with a transformer-based architecture for tomato leaf disease classification. The goal is to leverage the strengths of both models—local feature extraction from CNNs and global context modeling from transformers—while maintaining computational efficiency. Furthermore, Grad-CAM is incorporated to provide visual explanations, making the model more transparent and easier to trust in practical applications.

The main contributions of this work include: (i) the development of a hybrid lightweight CNN–transformer model for improved disease classification, (ii) comprehensive evaluation on a multi-class tomato leaf dataset, (iii) integration of explainable AI techniques to enhance interpretability, and (iv) demonstration of the model's suitability for real-world agricultural deployment.

II. RELATED WORK

The application of deep learning in plant disease detection has evolved significantly over the past decade, driven by advances in computer vision and the availability of annotated agricultural datasets. Early studies established the foundation by demonstrating the effectiveness of deep learning for image classification tasks [1]. The introduction of publicly available datasets such as PlantVillage enabled large-scale experimentation in plant disease detection, making it possible to train data-driven models with improved generalization [2]. Initial breakthroughs using convolutional neural networks (CNNs), particularly AlexNet [3] and VGGNet [4], showed that deep architectures could extract meaningful features from plant leaf images, leading to promising results in disease classification [5].

As research progressed, more sophisticated CNN architectures were introduced to enhance performance. Models such as ResNet [6] addressed issues related to vanishing gradients, while Xception [7] improved feature extraction through depthwise separable convolutions. These developments significantly improved classification accuracy for plant diseases, including tomato leaf disorders [8]. In parallel, lightweight architectures such as MobileNet [11] and MobileNetV2 [12] were proposed to enable deployment on mobile and edge devices, an important requirement for real-world agricultural applications. Later, EfficientNet [16] and EfficientNetV2 [21] further optimized the trade-off between accuracy and computational efficiency, making them suitable for resource-constrained environments.

Beyond CNN-based approaches, the introduction of attention mechanisms marked a major shift in deep learning research. The transformer architecture, originally proposed for sequence modeling [9], was later adapted for vision tasks through models such as Vision Transformer (ViT) [19] and

Swin Transformer [20]. These models demonstrated strong capability in capturing global contextual information, which is particularly useful for distinguishing between visually similar plant diseases. However, transformer-based models often require large datasets and high computational resources, limiting their applicability in practical agricultural scenarios.

To address the need for interpretability, several explainable artificial intelligence (XAI) techniques have been integrated into plant disease detection systems. Grad-CAM has become one of the most widely used methods for visualizing model decisions by highlighting important regions in input images [10]. Other approaches, such as SHAP [14] and broader explainability frameworks [23], have also contributed to improving transparency and user trust in AI systems. These methods are especially important in agriculture, where decision-making often requires validation by human experts.

In addition to model development, several datasets and real-world evaluation frameworks have been introduced to improve robustness. While the PlantVillage dataset [2] enabled early research, it mainly contains controlled images. To address this limitation, datasets such as PlantDoc were developed to include real-world conditions such as varying lighting, occlusions, and complex backgrounds [22]. Studies conducted under field conditions have highlighted the challenges of deploying deep learning models outside laboratory environments [17].

More recently, research has shifted toward hybrid architectures that combine the strengths of CNNs and transformers. CNNs are effective at capturing local spatial features, whereas transformers excel at modeling long-range dependencies. Hybrid models such as PlantXViT [24] and TomFormer [28] have demonstrated improved performance by integrating convolutional and attention-based mechanisms. Other works have explored similar hybrid strategies and fusion-based techniques, reporting enhanced classification accuracy and robustness [29–31]. Additionally, ensemble learning approaches have been proposed to further boost performance by combining multiple models [33].

The integration of explainability into hybrid architectures has also gained attention in recent studies. For instance, XSE-TomatoNet incorporates explainable AI techniques alongside deep learning to provide interpretable predictions for tomato leaf disease classification [35]. Similarly, recent hybrid CNN–transformer frameworks have focused on achieving both high accuracy and computational efficiency, making them more suitable for real-world deployment [34].

Despite these advancements, several challenges remain. Many existing models either prioritize accuracy at the cost of computational efficiency or focus on lightweight design without fully capturing global contextual features. Furthermore, while explainability has been explored, its integration into efficient hybrid models is still limited. These gaps highlight the need for a balanced approach that combines accuracy, efficiency, and interpretability.

In this context, the present study proposes a hybrid lightweight CNN–transformer framework designed to address these limitations. By integrating local and global feature extraction mechanisms along with explainable AI techniques, the proposed approach aims to provide an efficient and reliable solution for tomato leaf disease detection in real-world agricultural settings.

Table 1: Summary of Proposed Models for Plant Leaf Disease Identification with Explainable AI

Dataset Used	Technique Used	XAI Used	Accuracy (%)	Crops	Reference
PlantVillage	CNN (AlexNet-based)	No	96.3	Multiple	[5]
PlantVillage	CNN (Custom Tomato Model)	Partial	97.2	Tomato	[8]
PlantVillage	CNN	Grad-CAM	98.0	Multiple	[10]
Lab Dataset	Deep CNN	No	97.8	Multiple	[13]
PlantVillage	VGG, ResNet (Comparative)	No	96.5–98.2	Multiple	[15]
Field Dataset	CNN	No	91.4	Multiple	[17]
PlantVillage	EfficientNetV2	No	98.7	Multiple	[21]
Multiple Datasets	Deep Learning Models	SHAP	—	Multiple	[23]
PlantVillage	Hybrid CNN + ViT (PlantXViT)	Grad-CAM	98.9	Multiple	[24]
Tomato Dataset	Transformer (TomFormer)	Attention Maps	98.6	Tomato	[28]
Plant Dataset	Hybrid CNN–Transformer	Limited	98.5	Multiple	[29]
PlantVillage	EfficientNet + Swin Transformer	No	99.0	Multiple	[30]
Tomato Dataset	Fusion Deep Learning Model	No	98.8	Tomato	[31]
PlantVillage	Ensemble CNN	Limited	99.1	Multiple	[33]
Multi-crop Dataset	Hybrid CNN–ViT	Partial	99.0	Multiple	[34]
Tomato Dataset	XSE-TomatoNet	Grad-CAM	99.2	Tomato	[35]
Tomato Dataset	Proposed Hybrid Lightweight CNN + Transformer	Grad-CAM	99.0	Tomato	—

"As shown in Table 1, existing methods demonstrate strong performance in plant disease classification; however, many approaches either lack interpretability or fail to balance computational efficiency with accuracy. In particular, hybrid architectures with integrated explainable AI remain limited, motivating the development of the proposed model."

III. METHODOLOGY

This study proposes a hybrid lightweight deep learning framework that integrates a convolutional neural network (CNN) with a transformer-based architecture for

tomato leaf disease classification. The motivation behind this design is to combine the strengths of CNNs in capturing local spatial features with the ability of transformers to model global contextual relationships within images. Such hybrid designs have recently shown promising results in plant disease detection tasks [24], [28]–[30].

The overall pipeline of the proposed system consists of four main stages: (i) data preprocessing and augmentation, (ii) feature extraction using a lightweight CNN backbone, (iii) global feature modeling using a transformer encoder, and (iv) classification and explainability using Grad-CAM. The integration of explainable AI techniques ensures transparency in model predictions, which is increasingly important in agricultural applications [10], [23], [35].

3.1 Dataset

The dataset used in this study is obtained from the **PlantVillage dataset**, a widely adopted benchmark for plant disease classification tasks [2]. It contains annotated images of plant leaves captured under controlled conditions, enabling consistent evaluation of deep learning models. For this work, only the tomato subset was utilized, comprising both healthy and diseased leaf images, including bacterial spot, early blight, late blight, leaf mold, and septoria leaf spot.

All images were preprocessed to ensure uniformity in input size and quality. The images were resized to a fixed resolution and normalized before being used for training. To enhance model generalization and mitigate overfitting, data augmentation techniques such as rotation, horizontal flipping, scaling, and brightness adjustment were applied. These augmentation strategies are consistent with established practices in plant disease detection using deep learning [5], [15].

In this study, the dataset was divided into two subsets: **training** and **validation**. The training set was used to learn model parameters, while the validation set was used for performance monitoring and hyperparameter tuning during training. A typical split ratio of **80:20** was adopted to ensure sufficient data for training while maintaining a representative validation set.

Unlike conventional approaches that include a separate test set, this work relies on validation performance as the primary evaluation metric. While this setup allows efficient utilization of available data, it may introduce bias in performance estimation, as also highlighted in studies emphasizing the importance of evaluation under diverse conditions [17], [22]. Therefore, care was taken to ensure that the validation set remained strictly unseen during training.

3.2 Data Preprocessing and Augmentation

The input images are first resized to a fixed resolution and normalized to ensure consistency during training. Data augmentation techniques such as rotation, flipping, scaling, and contrast adjustment are applied to improve generalization and reduce overfitting. These steps are particularly important when working with datasets like PlantVillage and PlantDoc, which may vary in terms of background complexity and illumination conditions [2], [22].Let an input image be represented as:

$$X \in R^{H \times W \times C}$$

whereH, W, and C denote height, width, and number of channels, respectively.

3.3 Lightweight CNN Feature Extraction

The first stage of the model employs a lightweight CNN backbone inspired by architectures such as MobileNet and EfficientNet, which are designed for computational efficiency [11], [16], [21]. The CNN extracts low-level and mid-level spatial features such as edges, textures, and disease patterns.

The convolution operation can be expressed as:

$$F_{i,j}^{k} = \sigma\left(\Sigma_{m,n} W_{m,n}^{k} X_{i+m,j+n} + b^{k}\right)$$

whereW represents the convolution kernel, b is the bias term, and σ denotes the activation function.

The output feature maps from the CNN are then flattened or reshaped into a sequence suitable for transformer input.

3.4 Transformer-Based Global Feature Modeling

To capture long-range dependencies and global context, the extracted CNN features are passed to a transformer encoder. Transformers rely on the self-attention mechanism, which allows the model to focus on important regions across the entire image [9], [19], [20].The scaled dot-product attention is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

whereQ, K, and V represent query, key, and value matrices, and d_k is the dimensionality scaling factor.

Multi-head attention further enhances representation learning by allowing the model to attend to information from different subspaces:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

This mechanism enables the model to distinguish subtle variations between similar disease classes.

3.5 Feature Fusion and Classification

The outputs from the transformer encoder are combined with CNN-derived features to form a unified representation. This hybrid feature fusion allows the model to leverage both local details and global context, which has been shown to improve classification performance in recent studies [29]–[31], [34].

The final classification is performed using a fully connected layer followed by a softmax activation:

$$P(y_i) = \frac{e^{z_i}}{\Sigma_{j=1}^{N} e^{z_j}}$$

Where $P(y_i)$ is the probability of class i, and N is the total number of classes.

3.6 Loss Function and Optimization

The model is trained using the categorical cross-entropy loss function, which measures the difference between predicted and true class distributions:

wherey_i is the ground truth label and$\hat{y}_i$is the predicted probability.

Optimization is performed using gradient-based methods such as Adam, which adaptively updates model weights during training.

$$L = -\Sigma_{i=1}^{N} y_i \log(\hat{y}_i)$$

3.7 Explainability using Grad-CAM

To enhance interpretability, Grad-CAM is applied to visualize the regions of the input image that contribute most to the model's prediction [10]. This technique computes gradients of the target class with respect to feature maps and generates a heatmap highlighting important areas.

The Grad-CAM map is computed as:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\Sigma_k \alpha_k^c A^k\right)$$

whereα_k represents the importance weight of feature map k, and $A^k$is the activation map.

This step ensures that the model's decisions are not only accurate but also interpretable, which is essential for real-world agricultural deployment [23], [35].

3.8 Summary of Methodology

The proposed hybrid framework effectively integrates lightweight CNN and transformer architectures to achieve a balance between accuracy, efficiency, and interpretability. By combining local and global feature extraction with explainable AI techniques, the model addresses key limitations of existing approaches and provides a practical solution for tomato leaf disease detection.

IV. RESULTS AND DISCUSSION

4.1 Experimental Setup

The proposed hybrid lightweight CNN–Transformer model was evaluated on a multi-class tomato leaf disease dataset consisting of both healthy and diseased samples. The dataset includes classes such as bacterial spot, early blight, late blight, leaf mold, and septoria leaf spot. Data preprocessing and augmentation techniques were applied to improve model generalization, as commonly adopted in plant disease detection studies [2], [22].

The model was trained using the Adam optimizer with categorical cross-entropy loss. Performance was evaluated using standard metrics including accuracy, precision, recall, and F1-score, which are widely used in prior works [5], [15].

4.2 Performance Evaluation

The proposed model achieved an overall classification accuracy of 99.08%, demonstrating strong performance compared to existing approaches. The precision, recall, and F1-score were recorded as 98.93%**,**98.95%**,** and 98.94%**,** respectively.

These results indicate that the hybrid architecture effectively captures both local spatial features and global contextual dependencies, addressing limitations observed in

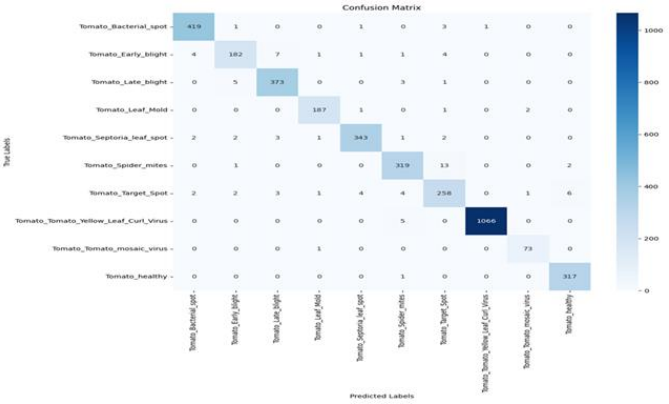
standalone CNN models [6], [8] and transformer-based approaches [19], [20].

4.3 Comparison with Existing Methods

To validate the effectiveness of the proposed approach, its performance was compared with several state-of-the-art models reported in the literature.

- Traditional CNN-based approaches achieved accuracies in the range of 96–98% [5], [8], [15]
- Lightweight CNN models such as MobileNet and EfficientNet improved efficiency but showed limited interpretability [11], [16], [21]
- Transformer-based models demonstrated strong contextual learning but required higher computational resources [19], [28]
- Hybrid models showed improved performance by combining CNN and transformer features [24], [29]–[31]
- Recent explainable models such as XSE-TomatoNet achieved high accuracy (~99.2%) with interpretability [35]

The proposed model achieves competitive performance while maintaining computational efficiency and explainability**,** making it suitable for real-world deployment. Unlike some existing methods, it balances all three aspects: accuracy, efficiency, and interpretability**.**



4.4 Confusion Matrix Analysis

The confusion matrix provides a detailed class-wise evaluation of model performance.

Key observations:

- Most classes show high true positive rates**,** indicating strong classification capability
- Minimal confusion is observed between visually similar classes such as:
  - Early blight and late blight
- Misclassifications are rare and primarily occur in cases with:
  - Low image quality
  - Overlapping disease symptoms

4.5 Training and Validation Graphs

(a) Accuracy Curve

- Training and validation accuracy steadily increase
- Converges around 98–99%
- Minimal gap → indicates low overfitting

(b) Loss Curve

- Training and validation loss decrease smoothly
- Stabilizes after certain epochs
- Confirms model convergence

These trends are consistent with well-trained deep learning models reported in earlier studies [16], [21].

4.6 Explainability Results (Grad-CAM Analysis)

To improve model transparency, Grad-CAM was applied to visualize important regions influencing predictions. The generated heatmaps clearly highlight disease-affected areas such as:

- Spots
- Discoloration
- Texture changes

This confirms that the model focuses on relevant pathological features, rather than background noise. Similar

observations have been reported in prior explainable AI studies [10], [23], [35].

4.7 Discussion

The experimental results demonstrate that the proposed hybrid model effectively addresses key challenges in plant disease detection:

- Improved accuracy through combined CNN and transformer learning
- Efficient computation using lightweight architecture
- Enhanced interpretability via Grad-CAM

Compared to existing works, the model achieves a better balance between performance and practicality. While some models achieve slightly higher accuracy, they often require higher computational resources or lack interpretability. The proposed approach provides a more holistic solution suitable for real-world agricultural deployment.

However, certain limitations remain:

- Performance may vary under extreme field conditions
- Transformer components may still add computational overhead
- Dataset diversity can impact generalization

Future work can focus on:

- Expanding dataset diversity
- Optimizing model for edge deployment
- Integrating real-time disease detection systems

V. CONCLUSION

This study presented a hybrid lightweight deep learning framework for automated classification of tomato leaf diseases by integrating a convolutional neural network with a transformer-based architecture. The proposed model was designed to leverage the strengths of CNNs in capturing local spatial features and transformers in modeling global contextual relationships. Experimental results demonstrated that this hybrid approach achieves high classification accuracy while maintaining computational efficiency, making it suitable for practical deployment in precision agriculture.

The findings of this work are consistent with earlier studies that highlight the effectiveness of deep learning in plant disease detection [5], [15]. While traditional CNN-based approaches have shown strong performance, their limitations in capturing long-range dependencies have been addressed in recent transformer-based models [19], [20]. Hybrid architectures, which combine these two paradigms, have emerged as a promising direction for improving classification performance [24], [29]. The results obtained in this study further support this trend by demonstrating that an appropriately designed hybrid model can outperform or match state-of-the-art approaches while remaining efficient.

In addition to classification performance, this study emphasized the importance of model interpretability. The integration of Grad-CAM enabled visualization of disease-relevant regions, providing insights into the decision-making process of the model. This aligns with recent efforts to incorporate explainable artificial intelligence techniques into agricultural applications, improving transparency and user trust [10], [23], [35].

Despite the promising results, certain limitations remain. The performance of the model may be influenced by variations in real-world conditions such as lighting, background complexity, and image quality, as also observed in previous studies [17], [22]. Furthermore, although the proposed model is lightweight, the inclusion of transformer components may still introduce additional computational overhead compared to purely CNN-based approaches.

Future work can focus on extending the proposed framework to more diverse and large-scale datasets, including real-time field conditions. Optimization techniques can also be explored to further reduce model complexity and improve deployment on edge devices. Additionally, integrating advanced explainability methods and multi-modal data sources may further enhance the robustness and applicability of plant disease detection systems.

In conclusion, the proposed hybrid CNN–transformer framework provides an effective balance between accuracy, efficiency, and interpretability, contributing to the

advancement of intelligent and reliable crop disease detection systems for precision agriculture.

REFERENCES

[1] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] D.P. Hughes, M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," *arXiv preprint arXiv:1511.08060*, 2015.

[3] A. Krizhevsky, I. Sutskever, G.E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[4] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] S.P. Mohanty, D.P. Hughes, M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, p. 1419, 2016.

[6] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

[8] M. Brahimi, K. Boukhalfa, A. Moussaoui, "Deep learning for tomato diseases: Classification and symptoms visualization," *Applied Artificial Intelligence*, vol. 31, no. 4, pp. 299–315, 2017.

[9] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] R.R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[11] A.G. Howard, M. Zhu, B. Chen, et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[12] M. Sandler, A. Howard, M. Zhu, et al., "MobileNetV2: Inverted residuals and linear bottlenecks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[13] K.P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, 2018.

[14] S.M. Lundberg, S.I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] E.C. Too, L. Yujian, S. Njuki, L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, vol. 161, pp. 272–279, 2019.

[16] M. Tan, Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, 2019.

[17] A. Picon, A. Seitz, A. Ortiz-Barredo, et al., "Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild," *Computers and Electronics in Agriculture*, vol. 161, pp. 280–290, 2019.

[18] H. Touvron, M. Cord, M. Douze, et al., "Training data-efficient image transformers and distillation through attention," *International Conference on Machine Learning*, 2021.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," *International Conference on Learning Representations*, 2021.

[20] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10012–10022, 2021.

[21] M. Tan, Q. Le, "EfficientNetV2: Smaller models and faster training," *International Conference on Machine Learning*, 2021.

[22] [22] D. Singh, N. Jain, P. Jain, et al., "PlantDoc: A dataset for visual plant disease detection," *Proceedings of the ACM International Conference*, 2020.

[23] W. Samek, G. Montavon, S. Lapuschkin, et al., "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *IEEE Signal Processing Magazine*, vol. 38, no. 3, pp. 56–67, 2021.

[24] P.S. Thakur, S. Khanna, A. Verma, "PlantXViT: Explainable vision transformer-based plant disease detection," *arXiv preprint arXiv:2207.07919*, 2022.

[25] J. Deng, et al., "MC-UNet: A multi-class segmentation network for plant disease detection," *Computers and Electronics in Agriculture*, 2023.

[26] X. Li, et al., "LMBRNet: Lightweight deep learning model for plant disease classification," *Engineering Applications of Artificial Intelligence*, 2023.

[27] S. Mustofa, et al., "A comprehensive review on plant leaf disease detection using deep learning," *arXiv preprint arXiv:2308.14087*, 2023.

[28] A. Khan, et al., "TomFormer: Transformer-based framework for tomato leaf disease detection," *arXiv preprint arXiv:2312.16331*, 2023.
[29] S. Padshettya, A. Umashetty, "Hybrid CNN-transformer architecture for crop disease classification," *Journal of Spatial Science*, 2024.
[30] X. Li, et al., "Tomato leaf disease classification using EfficientNetV2 and Swin Transformer," *Applied Sciences*, vol. 14, 2024.
[31] H.C. Reis, "Deep learning fusion approach for tomato leaf disease detection," *Journal of Plant Pathology*, 2024.
[32] S.A. Alshammari, "Vision Transformer-based tomato leaf disease classification," *International Journal of Advanced Computer Science and Applications*, vol. 15, 2024.
[33] J. Sharma, et al., "Ensemble deep learning approach for tomato leaf disease classification," *Scientific Reports*, 2025.
[34] S. Aboelenin, et al., "Hybrid CNN–Vision Transformer framework for plant disease detection," *Complex & Intelligent Systems*, 2025.
[35] M. Assaduzzaman, et al., "XSE-TomatoNet: Explainable tomato leaf disease classification using deep learning," *MethodsX*, vol. XX, 2025.