Feature-Level Fusion of CNN and Vision Transformer for Tomato Leaf Disease Identification

Afnan Ahmed[1], Sadiq Ali[1]
[1]Department of Electrical Engineering University of Engineering Technology Peshawar, Pakistan
*Correspondence: afnanahmed.eec@uetpeshawar.edu.pk, sadiqali@uetpeshawar.edu.pk

Citation | Ahmed. A, Ali. S, "Feature-Level Fusion of CNN and Vision Transformer for Tomato Leaf Disease Identification", IJIST, Vol. 07 Special Issue. pp 38-49, May 2025
Received | April 03, 2025 **Revised |** April 30, 2025 **Accepted |** May 02, 2025 **Published |** May 04, 2025. DOI:https://doi.org/10.33411/IJIST/1307

Tomato leaf diseases pose a serious threat to crop yield and quality, necessitating timely and accurate detection for effective management. Traditional visual inspection methods are subjective, labor-intensive, and inefficient, highlighting the need for automated solutions. This study explores the use of transfer learning and fine-tuning of deep learning models, ResNet-50 and Vision Transformers (ViT), for tomato leaf disease detection. A novel hybrid model integrating ResNet-50 and ViT through feature-level fusion is proposed to enhance classification accuracy. While ResNet-50 and ViT achieved accuracies of 95.20% and 98%, respectively, the hybrid model outperformed both with 99.07% accuracy. These results demonstrate the effectiveness and scalability of the hybrid model for early disease detection, offering a promising solution to enhance crop health and agricultural productivity.

Keywords: Leaf disease detection, Hybrid Model, Deep Learning Models

Introduction:

Globally, among the top three broadly traded vegetables, one of them is Tomatoes, which play a crucial part in the international vegetable market [1]. As part of the daily increase in the demand for tomatoes, their production on the worldwide level plus the area of cultivation keeps growing. At the same time, the preferences of consumers are shifting towards eco-friendly and high-quality products, which results in the increased need to improve the standards of food quality. However, the production and quality of tomatoes are often compromised due to various diseases that lead to a significant reduction in yield and, eventually, economic losses for poor farmers [2]. Plant diseases and insect damage are among the leading causes of agricultural losses globally. Estimations suggest that the losses in annual production are mostly due to pests and diseases, which have been substantial since the start of the 21st century [3]. On the other hand, plant diseases alone specifically contribute to the annual losses of 15% to 17% (approximately) of the total crop, which is a highly alarming figure! [4]. An estimated amount of 68% of the total annual production of the crop is lost due to factors such as pests, weeds, and plant leaf diseases [5], which causes a major economic setback. To address these challenges, there is a requirement to integrate advanced crop protection and enhancement strategies, utilizing and maximizing the latest global practices and emerging technologies. To ensure that the yield and quality of tomatoes are significant not only for food security but also for global economic trade. Conventionally, the identification of tomato leaf disease relied on manual visual inspection done visually by field workers, a method prone to subjectivity, inefficiency, and low accuracy. Considering these limitations, it is essential to develop efficient and automated methods for detecting tomato leaf diseases and pests using modern technology. In response to these challenges, we propose:

- To develop a robust tomato leaf disease detection model that generalizes well to real-world conditions, we enhanced the existing dataset using extensive data augmentation techniques, including variations in lighting, orientation, and background noise. This improved diversity allows the model to perform robustly under practical agricultural scenarios, significantly boosting the reliability of automated disease detection for effective crop health management.
- Existing tomato leaf disease datasets often lack diversity, limiting model generalization and real-world accuracy. By creating a more diverse and representative dataset, we aim to improve the model's adaptability across different environmental conditions.
- Traditional CNNs primarily capture local features, while Vision Transformers (ViTs) focus on global dependencies, limiting their standalone effectiveness in fine-grained disease detection. To address this, we fine-tuned pre-trained CNN and ViT models using transfer learning and proposed a hybrid CNN-transformer model that combines local and global feature extraction. This fusion leverages the strengths of both architectures, significantly enhancing classification performance for tomato leaf disease detection.
- By integrating CNN and Transformer models and augmenting the dataset comprehensively, this study introduces a novel approach that significantly improves fine-grained disease detection accuracy, ensuring better real-world deployment in agricultural practices.

Related Work:

Many studies have explored plant disease detection, particularly in tomato leaves, but limitations persist. Traditional methods rely on principal component analysis, using a single sample leaf as a reference, and texture-based segmentation techniques [6][7], which struggle in real-time scenarios [8][9]. To address these challenges, [10] introduced PLPNet, which tackles intraclass variability and similarity, key factors in disease classification. Object detection techniques were emphasized to improve accuracy, particularly in cases where soil backgrounds obscure infected leaf edges. Building on this, [11] proposed TomatoDet, integrating Swin-

DDETR's self-focus mechanism, Meta-ACON launch, and an improved bidirectional weighted feature pyramid network (IBiFPN) to enhance small-target disease identification and reduce false detections. Transformer-based approaches were further explored in [12] with the NanoSegmenter model, incorporating lightweight techniques such as quantization and sparse attention to optimize efficiency. It achieved a precision of 0.98, recall of 0.97, and more of 0.95, with an inference speed of 37 FPS, making it viable for real-time agricultural applications. In [13], traditional image classification models were compared with the YOLO object detection framework, where optimized feature layers and attention mechanisms improved real-time decision-making for early pest detection and crop loss reduction. Beyond tomato crops, deep learning has been applied to plant disease detection more broadly.

In [14], transfer learning was used to train a deep CNN on cassava disease images, achieving up to 98% accuracy for certain diseases. A comparison of machine learning models found that SVM outperformed others in four out of six disease categories. Lastly, [15] highlighted the importance of deep learning in food security, training a deep CNN on a public dataset of 14 crop species and 26 diseases. While the model achieved high accuracy, performance dropped under varying environmental conditions, underscoring the need for diverse training data. Despite limitations, increasing smartphone accessibility presents opportunities for AI-driven disease detection to aid farmers and enhance crop management.

Proposed Methodology:

The proposed framework comprises four key stages: Data Acquisition (for collecting labeled disease images), Pre-processing (to standardize and normalize inputs), Augmentation (to increase dataset diversity and generalization), and Classification (using deep learning models). The complete workflow is illustrated in Figure 1.

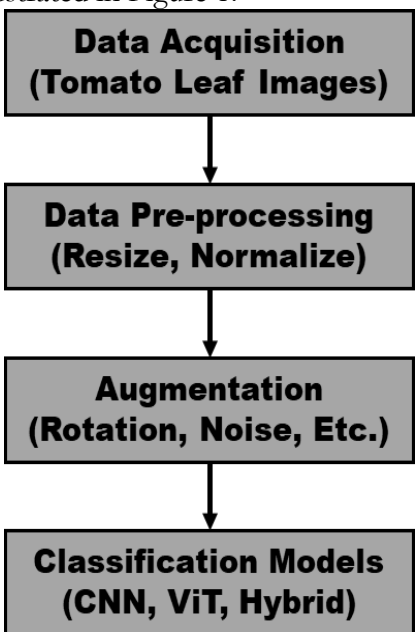


Figure 1. Proposed Methodology

Data Acquisition:

The dataset of tomato leaf diseases consists of images collected from public sources and agricultural research databases, reflecting variations in leaf conditions due to factors like climate, soil, and farming practices. The dataset includes ten disease classes: Early Blight, Late Blight, Septoria Leaf Spot, Bacterial Spot, Target Spot, Leaf Mold, Yellow Leaf Curl Virus, Mosaic Virus, Powdery Mildew, and Healthy Leaves. Key features include lesion texture, shape, color, and leaf structure, which are critical for accurate disease classification. This dataset facilitates early disease detection, enabling timely interventions to reduce yield losses.

The dataset's diversity strengthens machine learning models, making them scalable and adaptable for better crop management. Sample images are shown in Table 1. Although the images were collected from multiple sources and captured under varying lighting, backgrounds, and environmental conditions, this diversity was intentionally retained to enhance the model's generalization capability in real-world scenarios.

Data Pre-Processing:

To ensure the dataset's consistency and suitability for deep learning, several pre-processing techniques were applied. The images were systematically renamed based on their respective class, appending an incremental numeric identifier (e.g., "EarlyBlight_1," "EarlyBlight_2"). All images were resized uniformly to 224×224 pixels to maintain consistent input dimensions for deep learning models. In the final step, pixel values were normalized to a [0,1] range by dividing by 255, optimizing model performance during training and evaluation. The original images were generally of high quality. While no specific denoising filters were applied, basic noise inspection was conducted. Preserving natural image variations was a priority, and the resizing and normalization steps helped standardize input distributions for model stability.

Table 1. Sample Pictures of tomato Leaf Disease dataset

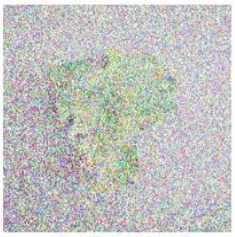
 Bacterial Spot	 Early Blight	 Healthy
 Late Blight	 Leaf Mold	 Septoria Leaf Spot
 Yellow Leave Curl	 Spider Mites	 Target Spot
 Tomato Mosaic Virus		

Augmentation:

To improve the robustness of the dataset, various data augmentation techniques were applied. These included random rotation (from -30 to 30 degrees) and horizontal flipping,

which increased viewpoint diversity, as well as brightness and contrast adjustments to simulate varying lighting conditions. Additionally, subtle variations and random cropping simulated zooming effects. Gaussian noise (mean 0, std 25) was also added to introduce minor noise, further enhancing dataset diversity and improving the model's adaptability to real-world conditions. Sample augmented images are shown in Table 2.

Table 2. Sample Pictures of Augmentation

 Random Rotation	 Brightness Adjustment	 Random Cropping
 Horizontal Flip	 Gaussian Noise	 Contrast Adjustment

Classification Model:

This study utilizes and implements a deep learning-based classification framework for the detection of Tomato Leaf Disease by the integration of ResNet-50, Vision Transformer (ViT-B_16), and a Feature-Level Fusion Hybrid Model. ResNet-50 (Figure 2) is a model that is fine-tuned to capture the patterns that are disease-specific, such as texture variations and discoloration, employing transfer learning to hold its general feature extraction while performing the processes to refine deeper layers for precise recognition of disease. The robustness of the model was enhanced with a custom classification head, along with dynamic learning rate scheduling, regularization, and data augmentation.

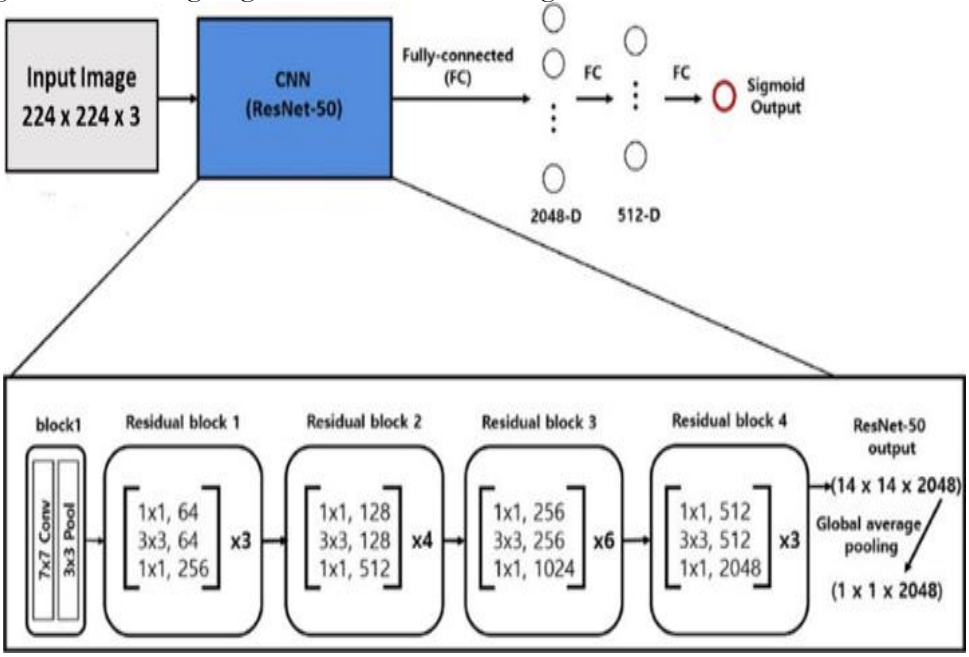


Figure 2. ResNet-50 Architecture

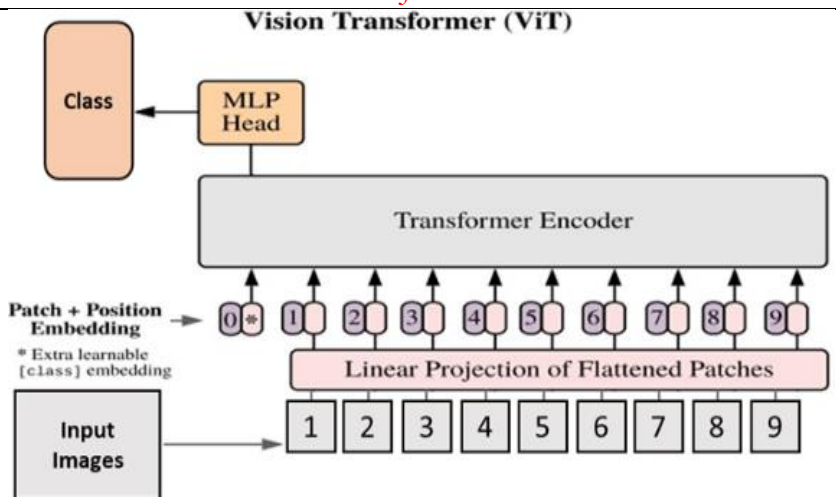


Figure 3. Vision Transformer Architecture

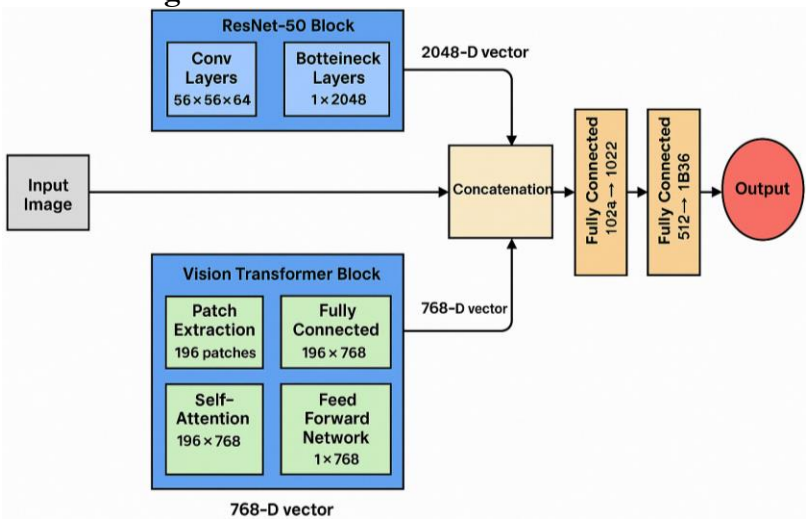


Figure 4. Feature-Level Fusion Hybrid Model Architecture

It also prevents overfitting, ensuring reliable detection of disease [16]. ViT-B_16 (Figure 3) uses a mechanism of self-attention to capture global dependencies and intricate disease features. Fine-tuned on the dataset of tomato leaves, ViT-B_16 can effectively distinguish between healthy and diseased leaves, enabling intervention at an early stage. Task-specific layers and dynamic hyperparameter tuning further improve the accuracy of classification, making the model highly adaptable to complex disease patterns [17].

To utilize the strengths of both architectures, a Feature-Level Fusion Hybrid Model (Figure 4) integrates ResNet-50 and ViT-B_16. ResNet-50 extracts fine-grained spatial details, while ViT-B_16 captures long-range dependencies that provide a comprehensive understanding of the symptoms of the disease. The fusion of feature representations through specialized layers enhances the accuracy of classification, which ensures robustness in the identification of disease for effective protection of the crop.

Experimental Setup:

The model was implemented using Kaggle's GPU environment for efficient training. The dataset was split into 70% for training, 15% for validation, and 15% for testing. Key hyperparameters included 15 epochs, a 0.0005 learning rate, and a batch size of 32. Images were resized to 224x224x3 for standardization. A custom learning rate scheduler optimized convergence, while overfitting detection determined epoch limits. Gamma adjustment and a step schedule-controlled learning rate decay. For multi-class classification, class-specific weighting and a hybrid loss function (Categorical Cross-Entropy) improved generalization.

The Adam optimizer, enhanced with dynamic learning rate adjustments and gradient smoothing, ensured stable and efficient training, leading to strong classification performance [18].

Result and Analysis:

For Tomato Leaf Disease Detection, all three models demonstrated consistent improvement over 15 epochs. ResNet-50 achieved a training accuracy increase from 61.23% to 97.03%, with validation accuracy rising from 81.87% to 94.67%, while training and validation losses decreased to 0.0075 and 0.0092, respectively (Figure 5). ViT-B_16 outperformed ResNet-50, with training accuracy improving from 65.74% to 99.63% and validation accuracy stabilizing at 97.60%, alongside a steady decline in training and validation losses to 0.0014 and 0.0026, respectively (Figure 6). The hybrid model exhibited the most robust performance, leveraging both architectures to enhance feature representation. It achieved a training accuracy increase from 78.66% to 99.97%, with validation accuracy stabilizing at 98.70%. Training and validation losses steadily declined to 0.0196 and 0.0516, respectively (Figure 7).

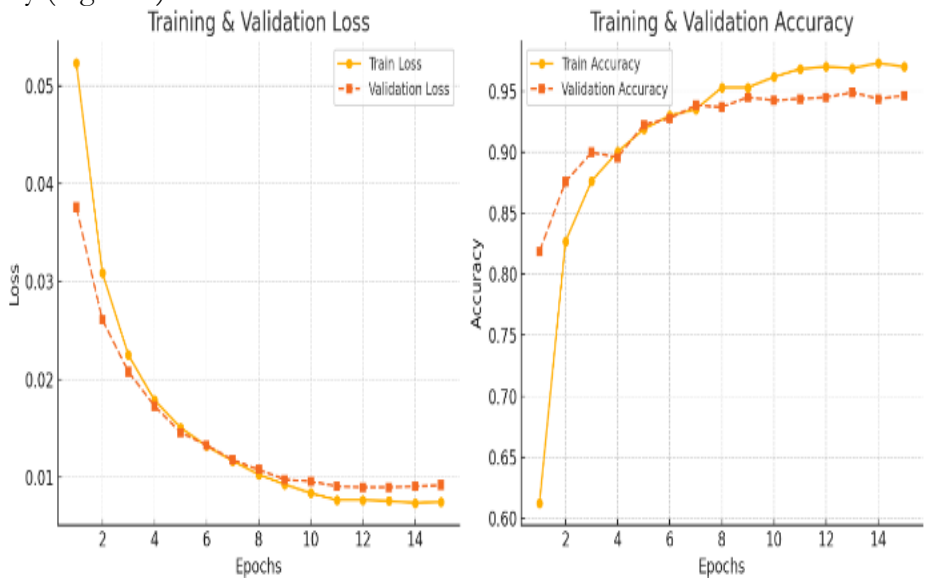


Figure 5. Training and Validation Accuracy and Loss ResNet-50

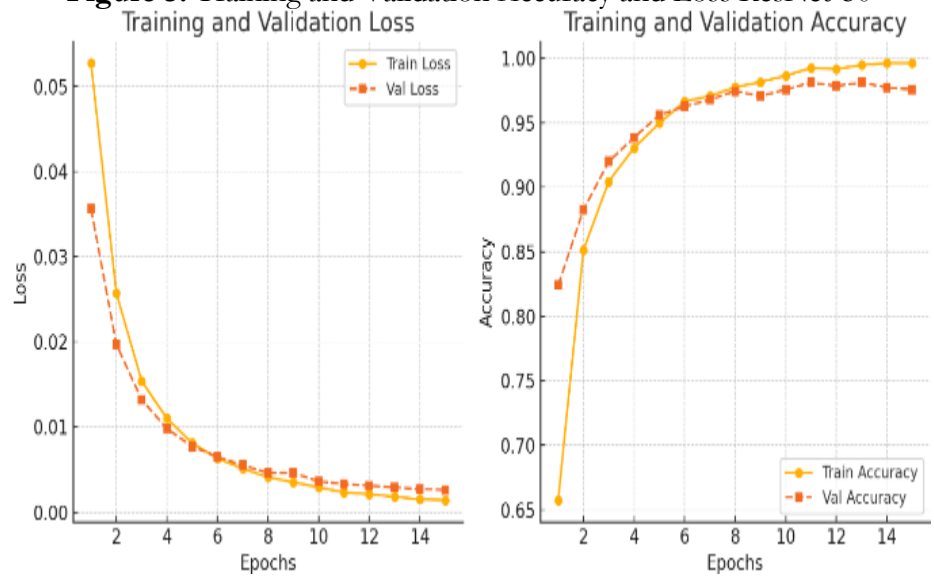


Figure 6. Training and Validation Accuracy and Loss Vision Transformer

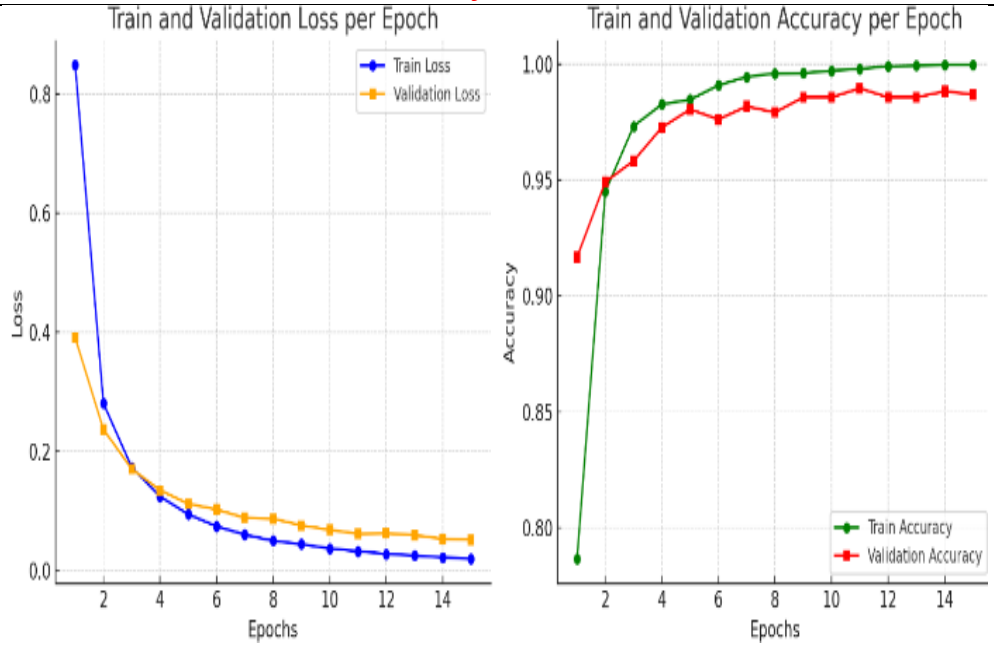


Figure 7. Training and Validation Accuracy and Loss Hybrid Model

Performance metrics in Table 3 further validate these findings. ResNet-50 achieved an accuracy of 95%, with recall, precision, and F1-score all at 95%, indicating reliable classification. ViT-B_16 demonstrated superior performance, achieving 98% across all evaluation metrics. The hybrid model outperformed both, achieving the highest accuracy of 99%, along with 99% recall, precision, and F1-score. It is also important to highlight that the dataset exhibited moderate class imbalance, with some disease classes represented by fewer samples. To address this, targeted data augmentation techniques were applied to increase the sample size of underrepresented classes. Additionally, all reported metrics are macro-averaged to ensure that the performance evaluation remains unbiased and reflective of all classes equally. This approach prevents the model from favoring majority classes and ensures robust and fair classification across the entire dataset.

Table 3. Macro-averaged classification performance

Model	Recall	Precision	F1-Score	Accuracy
Resnet-50	95%	95%	95%	95%
ViT B_16	98%	98%	98%	98%
Hybrid Model	99%	99%	99%	99%

Confusion Matrix:

For Tomato Leaf Disease Detection, ResNet-50 achieved 95.2% accuracy but showed minor misclassifications, particularly between Early Blight and Septoria Leaf Spot (Figure 8). ViT-B_16 improved accuracy to 98.0%, enhancing class differentiation, though slight confusion remained in closely related diseases (Figure 9). The Hybrid model successfully achieved the highest accuracy of 99.07%, with a classification that is near-perfect and has minimal misclassifications in diseases that are visually similar (Figure 10). The results highlighted the superior precision of the hybrid model, which makes it highly effective for the early detection of disease in precision agriculture.

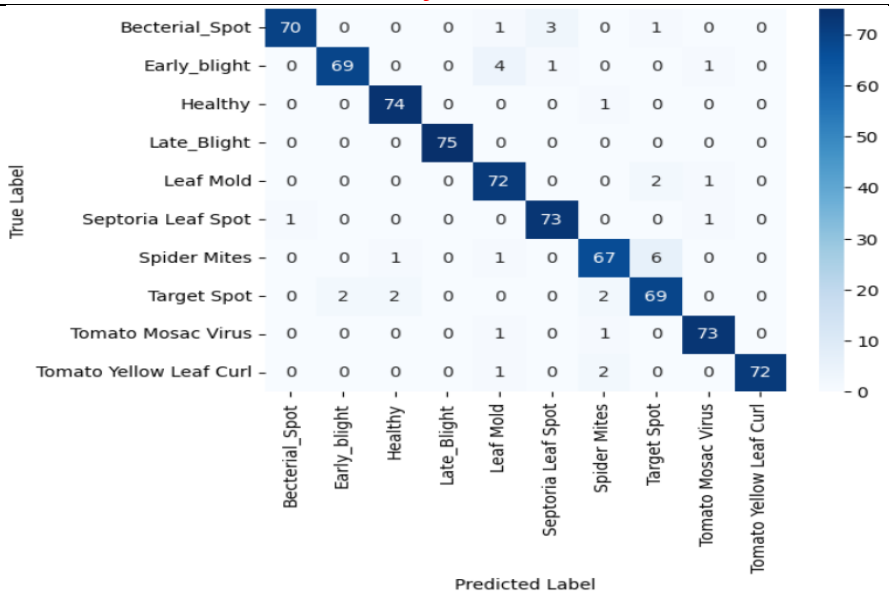


Figure 8. ResNet-50 Confusion Matrix

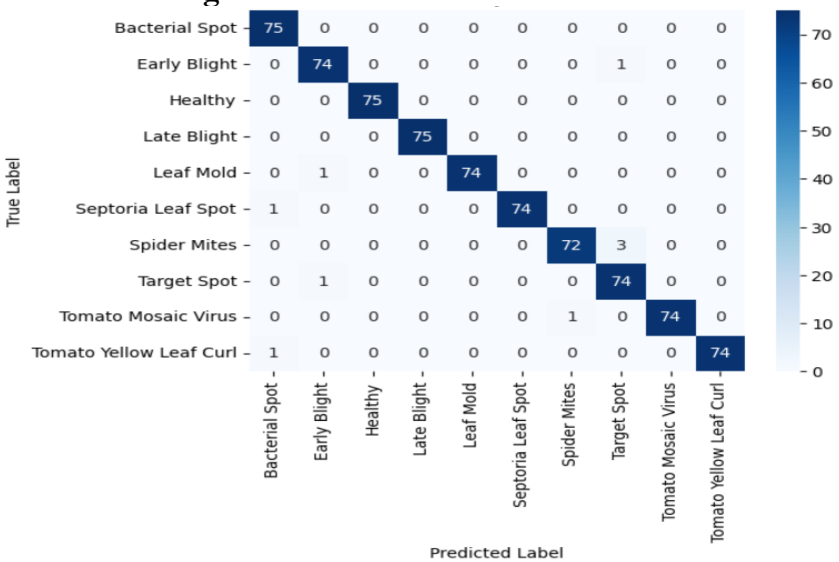


Figure 9. Vision Transformer Confusion Matrix

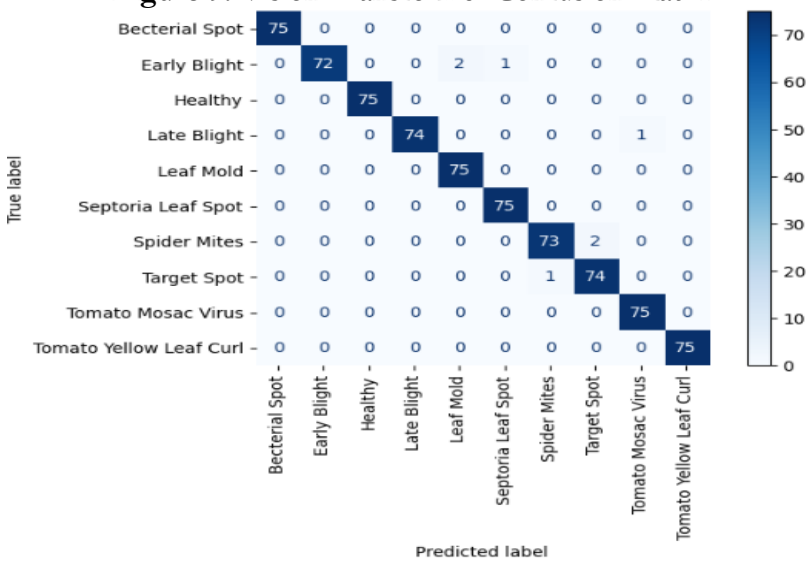


Figure 10. Hybrid Model Confusion Matrix

Discussion:

To evaluate the effectiveness of our proposed hybrid CNN-Transformer model, we compared its performance with that of established deep learning models, specifically ResNet-50 and Vision Transformer (ViT), as reported in prior studies. As summarized in Table 4, the ResNet-50 model in [19] achieved notably high accuracy (99.97%), recall (99.87%), and precision (99.86%). However, this model was trained without data augmentation, which may have resulted in overfitting and limited its generalizability to real-world conditions. Although it utilized the same ten-class tomato leaf disease dataset, the absence of image variability restricted its robustness. In contrast, the ViT model reported in [20] obtained a significantly lower accuracy of 90.99%, with a precision of 90.9% and recall of 89.3%. This reduction in performance suggests that ViT alone may struggle with detailed local feature extraction, which is critical for distinguishing subtle visual differences between disease categories. It is also important to highlight key differences in methodologies and preprocessing practices across these studies. The ResNet-50 and ViT models in the referenced works either lacked or applied minimal augmentation and normalization, leading to inconsistent image exposure and limited diversity in training data. In our study, we employed class-balanced data augmentation strategies, including brightness and contrast adjustment, flipping, rotation, and Gaussian noise, to simulate real-world variability. Additionally, all images were uniformly resized to 224×224 pixels and normalized to a [0,1] scale, ensuring consistency and improved training convergence. These methodological improvements contributed significantly to the enhanced robustness and generalization capability of our hybrid model. Our proposed hybrid model addresses the individual limitations of CNNs and Transformers by integrating local and global feature extraction capabilities. It achieved a classification accuracy of 99.07%, along with balanced precision, recall, and F1-score values of 99%, as shown in Table 4. These results indicate that our model not only matches the high accuracy of CNNs but also enhances robustness through better generalization a benefit derived from combining architecture types and applying comprehensive data augmentation. Overall, the comparative analysis highlights the strength of our hybrid approach in achieving high accuracy while maintaining generalization across diverse input conditions. This makes it a promising solution for practical deployment in automated crop disease detection systems.

Table 4. Comparative Analysis

Model	Recall	Precision	F1-Score	Accuracy
[19]	99.87%	99.86%	99.88%	99.97%
[20]	89.3%	90.9%	90.7%	90.99%
Hybrid Model	99%	99%	99%	99.07%

Conclusion:

For Tomato Leaf Disease Detection, ResNet-50 achieved 95.2% accuracy but showed minor misclassifications, particularly between Early Blight and Septoria Leaf Spot (Figure 8). ViT-B_16 improved accuracy to 98.0%, enhancing class differentiation, though slight confusion remained in closely related diseases (Figure 9). The Hybrid model successfully achieved the highest accuracy of 99.07%, with a classification that is near-perfect and has minimal misclassifications in diseases that are visually similar (Figure 10). The results highlighted the superior precision of the hybrid model, which makes it highly effective for the early detection of disease in precision agriculture.

References:

[1] P. L. Tong Li, Jiaxin Cui, Wei Guo, Yingjun She, "The Influence of Organic and Inorganic Fertilizer Applications on Nitrogen Transformation and Yield in Greenhouse Tomato Cultivation with Surface and Drip Irrigation Techniques," Water, vol. 15, no. 20, p. 3546, 2023, doi: https://doi.org/10.3390/w15203546.

[2] Yuanhui Yu, "Research Progress of Crop Disease Image Recognition Based on

Wireless Network Communication and Deep Learning," Wirel. Commun. Mob. Comput., 2021, doi: https://doi.org/10.1155/2021/7577349.

[3] H. Durmus, E. O. Gunes, and M. Kirci, "Disease detection on the leaves of the tomato plants by using deep learning," 2017 6th Int. Conf. Agro-Geoinformatics, Agro-Geoinformatics 2017, Sep. 2017, doi: 10.1109/AGRO-GEOINFORMATICS.2017.8047016.

[4] E. D. I. Valenzuela, R. Baldovino, A. Bandala, "Pre-Harvest Factors Optimization Using Genetic Algorithm for Lettuce," J. Telecommun. Electron. Comput. Eng., vol. 10, no. 1–4, pp. 159–163, 2018, [Online]. Available: https://jtec.utem.edu.my/jtec/article/view/3610

[5] R. G. De Luna et al., "Identification of philippine herbal medicine plant leaf using artificial neural network," HNICEM 2017 - 9th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag., vol. 2018-January, pp. 1–8, Jul. 2017, doi: 10.1109/HNICEM.2017.8269470.

[6] I. C. Valenzuela et al., "Quality assessment of lettuce using artificial neural network," HNICEM 2017 - 9th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag., vol. 2018-January, pp. 1–5, Jul. 2017, doi: 10.1109/HNICEM.2017.8269506.

[7] J. B. U. Dimatira et al., "Application of fuzzy logic in recognition of tomato fruit maturity in smart farming," IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, pp. 2031–2035, Feb. 2017, doi: 10.1109/TENCON.2016.7848382.

[8] I. C. Valenzuela, R. G. Baldovino, A. A. Bandala, and E. P. Dadios, "Optimization of Photosynthetic Rate Parameters using Adaptive Neuro-Fuzzy Inference System (ANFIS)," 2017 Int. Conf. Comput. Appl. ICCA 2017, pp. 129–134, Oct. 2017, doi: 10.1109/COMAPP.2017.8079734.

[9] J. Shijie, J. Peiyi, H. Siping, and Sl. Haibo, "Automatic detection of tomato diseases and pests based on leaf images," Proc. - 2017 Chinese Autom. Congr. CAC 2017, vol. 2017-January, pp. 3507–3510, Dec. 2017, doi: 10.1109/CAC.2017.8243388.

[10] Y. H. Zhiwen Tang, Xinyu He, Guoxiong Zhou, Aibin Chen, Yanfeng Wang, Liujun Li, "A Precise Image-Based Tomato Leaf Disease Detection Approach Using PLPNet," Plant Phenomics, vol. 5, p. 0042, 2023, doi: https://doi.org/10.34133/plantphenomics.0042.

[11] Xuewei Wang & Jun Liu, "An efficient deep learning model for tomato disease detection," Plant Methods, vol. 20, no. 61, 2024, doi: https://doi.org/10.1186/s13007-024-01188-1.

[12] T. C. Yufei Liu, Yihong Song, Ran Ye, Siqi Zhu, Yiwen Huang, "High-Precision Tomato Disease Detection Using NanoSegmenter Based on Transformer and Lightweighting," Plants, vol. 12, no. 13, p. 2559, 2023, doi: https://doi.org/10.3390/plants12132559.

[13] X. W. Jun Liu, "Tomato Diseases and Pests Detection Based on Improved Yolo V3 Convolutional Neural Network," Front. Plant Sci, vol. 11, 2020, doi: https://doi.org/10.3389/fpls.2020.00898.

[14] A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, "Deep learning for image-based cassava disease detection," Front. Plant Sci., vol. 8, no. October, pp. 1–7, 2017, doi: 10.3389/fpls.2017.01852.

[15] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," Front. Plant Sci., vol. 7, no. September, Sep. 2016, doi: 10.3389/FPLS.2016.01419.

[16] I. Z. Mukti and D. Biswas, "Transfer Learning Based Plant Diseases Detection Using ResNet50," 2019 4th Int. Conf. Electr. Inf. Commun. Technol. EICT 2019, Dec. 2019,

doi: 10.1109/EICT48899.2019.9068805.

[17] A. Tabbakh and S. S. Barpanda, "A Deep Features Extraction Model Based on the Transfer Learning Model and Vision Transformer 'TLMViT' for Plant Disease Classification," IEEE Access, vol. 11, pp. 45377–45392, 2023, doi: 10.1109/ACCESS.2023.3273317.

[18] R. O. Ogundokun, R. Maskeliunas, S. Misra, and R. Damaševičius, "Improved CNN Based on Batch Normalization and Adam Optimizer," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 13381 LNCS, pp. 593–604, 2022, doi: 10.1007/978-3-031-10548-7_43.

[19] A. D. K. Muslih, "Tomato Leaf Diseases Classification using Convolutional Neural Networks with Transfer Learning Resnet-50," Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control, vol. 9, no. 2, 2024, doi: https://doi.org/10.22219/kinetik.v9i2.1939.

[20] M. J. S. Utpal Barman, Parismita Sarma, Mirzanur Rahman, Vaskar Deka, Swati Lahkar, Vaishali Sharma, "ViT-SmartAgri: Vision Transformer and Smartphone-Based Plant Disease Detection for Smart Agriculture," Agronomy, vol. 14, no. 2, p. 327, 2024, doi: https://doi.org/10.3390/agronomy14020327.

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.