International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Explainable vision transformers for real-time chili and onion leaf disease identification and diagnosis

Hamdadur Rahman [1], Hasan Md Imran [2,*], Amira Hossain [3], Md Ismail Hossain Siddiqui [4] and Anamul Haque Sakib [5]

[1] *Department of Management Information System, International American University, Los Angeles, CA 90010, USA.*
[2] *Department of Computer Science, The University of Alabama in Huntsville, Huntsville, AL 35899, USA.*
[3] *Department of Computer Science, Westcliff University, Irvine, CA 92614, USA.*
[4] *Department of Engineering/Industrial Management, Westcliff University, Irvine, CA 92614, USA.*
[5] *Department of Business Administration, International American University, Los Angeles, CA 90010, USA.*

International Journal of Science and Research Archive, 2025, 15(01), 1823-1833

Publication history: Received on 13 March 2025; revised on 22 April 2025; accepted on 24 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.1163

Abstract

Early identification of leaf diseases in chili and onion crops is crucial for maintaining agricultural productivity and reducing economic losses. This study proposes a transformer-based deep learning framework for the multi-class classification of common leaf diseases affecting chili and onion plants. It addresses challenges related to intra-class similarity, complex backgrounds, and variations in real-world imaging. We collected a curated dataset consisting of 13,989 high-resolution images—10,987 of chili leaves and 4,502 of onion leaves—from actual agricultural environments in Karnataka, India. This dataset covers nine disease classes, including Cercospora, purple blotch, Iris yellow spot virus, and powdery mildew. To enhance model generalization, we applied extensive preprocessing techniques, including resizing, normalization, augmentation, and noise injection. We evaluated four state-of-the-art transformer architectures: MaxViT, Swin Transformer, Hornet, and EfficientFormer. Among these, MaxViT achieved the highest performance, with classification accuracies of 95.75% on the onion dataset and 90.86% on the chili dataset, along with high F1 scores, Matthews Correlation Coefficient (MCC), and Precision-Recall Area Under Curve (PR-AUC) values. To enable practical use in the field, we developed a real-time web application using Django. This application allows users to upload leaf images and receive instant predictions, supplemented by Grad-CAM-based visual explanations. This integration of explainable AI (XAI) enhances transparency and builds trust among end-users, such as farmers and agronomists. The results highlight the effectiveness of transformer-based models for agricultural disease diagnosis and provide a scalable, interpretable, and deployable solution for precision farming.

Keywords: Plant disease; Vision transformer; Explainable AI; Chili leaves; Onion leaves; Agricultural monitoring

1 Introduction

Chili and onion are among the most widely cultivated vegetable crops across Asia, Africa, and Latin America. Globally, chili production exceeds 3.5 million metric tons annually, while onion ranks as the second most consumed vegetable crop, with over 16 million metric tons produced worldwide each year [1], [2]. These crops play a vital role in food security and the economy, serving as both staple ingredients and commercial commodities [3]. However, their productivity is frequently threatened by a variety of leaf diseases, including Cercospora, purple blotch, leaf blight, and powdery mildew. Studies estimate that leaf diseases can cause yield losses of up to 30–50% in chili and onion fields, particularly in tropical climates [2]. Such diseases not only reduce the quantity but also the quality of yield, leading to substantial economic setbacks and increasing dependence on chemical fungicides. Accurate identification of these leaf

* Corresponding author: Hasan Md Imran

diseases is therefore crucial for healthy crop management, minimizing pesticide overuse, and promoting sustainable agricultural practices [4].

Traditional plant leaf disease diagnosis relies on visual inspection by experts, which is often subjective, error-prone, and labor-intensive. These manual methods are difficult to scale across large agricultural fields, especially in regions with limited access to plant pathologists [5], [6]. Furthermore, many disease symptoms share overlapping visual traits, such as spots, blights, and discoloration, making them difficult to distinguish using conventional techniques. Variations in lighting, background, and leaf maturity further complicate accurate disease detection in real-world conditions [7].

Convolutional Neural Networks (CNNs) have emerged as a powerful tool for plant disease classification due to its ability to learn discriminative features directly from raw images. While CNNs have shown high accuracy in controlled environments, they often struggle with complex datasets where intra-class variation and inter-class similarity are high [8], [9]. Transformer-based architectures, such as MaxViT, Swin Transformer, Hornet, and EfficientFormer, offer a promising alternative by combining local and global feature extraction. These models are better equipped to handle subtle disease patterns and variable imaging conditions, making them suitable for field-level diagnosis.

In addition to model accuracy, real-world deployment requires solutions that are accessible and interpretable. To address this, we developed a lightweight web application that integrates the trained deep learning model for real-time inference. Users can upload leaf images through the browser and receive instant predictions, making the system practical for farmers, agronomists, and field workers. To further support decision-making and build trust, we incorporated XAI using Grad-CAM to highlight the specific regions of the leaf that influenced the model's prediction.

Previous studies in plant disease detection often rely on small, single-crop datasets, with limited focus on generalization and usability. Most do not include transformer-based architectures or support real-time use through deployment and explainability. There remains a gap in delivering high-performing, scalable, and interpretable systems that can serve diverse agricultural needs. This study addresses these gaps by evaluating multiple deep learning models across two real-world datasets of chili and onion leaves. Our key contributions are as follows:

- Conducted a comparative analysis of transformer-based models (MaxViT, Swin, Hornet, and EfficientFormer) for multi-class leaf disease classification.
- We have developed a unified pipeline integrating preprocessing, training and deployment for robust model performance.
- Deployed a web-based application that provides real-time prediction and user-friendly interaction.
- Incorporated Grad-CAM explainability to visualize model decisions, improving interpretability for non-expert users and field validation.

The rest of the paper is structured as follows: Section 2 presents related works on plant disease detection and highlights existing limitations. Section 3 describes the datasets, preprocessing techniques, and model architectures. Section 4 reports experimental results, including evaluation metrics and comparisons with state-of-the-art methods. Section 5 discusses findings, practical implications, and limitations. Finally, Section 6 concludes the paper and outlines future directions for research and deployment.

2 Related Works

The detection of leaf diseases in chili and onion crops using deep learning and machine learning techniques has gained significant attention due to their ability to improve disease management and optimize agricultural practices. Various studies have explored different approaches for disease detection, with notable improvements in accuracy and generalization.

For chili leaf disease detection, Pratap and Kumar [10] proposed a customized EfficientNetB4 model, fine-tuned to detect multiple chili leaf diseases. The model achieved an accuracy of 92%, outperforming other models like ResNet-50 and MobileNet-V2. However, the generalization ability to other crops was not evaluated. Similarly, a study by Muslim et al. [11] developed an Android-based disease detection system for rice and chili crops using MobileNet V1 and Sequential CNN models. Their model achieved a testing accuracy of 95%, demonstrating its potential for real-time use in the field. However, their study was limited to only a few chili diseases, limiting their broader applicability. In the realm of onion leaf disease detection.

McDonald et al. [12] explored aerial photography using UAVs equipped with a near-infrared (NIR) camera to assess Stemphylium leaf blight. While the study provided useful insights into disease severity through NDVI and other vegetative indices, no strong correlations between disease severity and the indices were found, suggesting limitations in the sensitivity of the indices for precise disease detection. On the other hand, Amondkar and Bhoite [13] employed

machine learning techniques to detect onion leaf diseases in Maharashtra, India. They achieved significant classification accuracy but encountered challenges with dataset diversity and the scalability of their approach.

In a broader context, Shao et al. [14] investigated the use of hyperspectral imaging to detect root rot in chili peppers, achieving a classification accuracy of 92.3% using a SPA-BP model. While hyperspectral imaging has shown promise for plant disease detection due to its ability to capture detailed spectral information, the study acknowledged the complexity of acquiring hyperspectral data and the need for specific wavelengths to achieve accurate results. This dependency on specialized equipment and the challenge of capturing data in real-world conditions make the approach less practical for large-scale, on-the-ground applications.

Major obstacles to broader adoption include the use of limited and homogenous datasets, environmental adaptation challenges, and inherent scalability restrictions. Additionally, many current models function as opaque "black box" systems, which lack the necessary explainability to elucidate decision-making processes. This study addresses these challenges by implementing robust approaches that increase dataset diversity and integrating XAI methods within a web-based framework, thereby enhancing identification accuracy and boosting user confidence in conservation progress.

3 Materials and Methods

Figure 1 depicts our proposed methodology for chili and onion leaf disease classification. The process starts with acquiring input images, which undergo preprocessing that includes resizing, flipping, brightness adjustment, and adding Gaussian noise for better generalization. These images are then processed through four transformer-based deep learning models for multi-class classification. To enhance transparency, Grad-CAM is used to highlight key areas that influence predictions. A real-time web application, Onichili: Leaf Disease Classifier, allows users to upload images, receive immediate predictions, and see visual interpretations through XAI overlays.

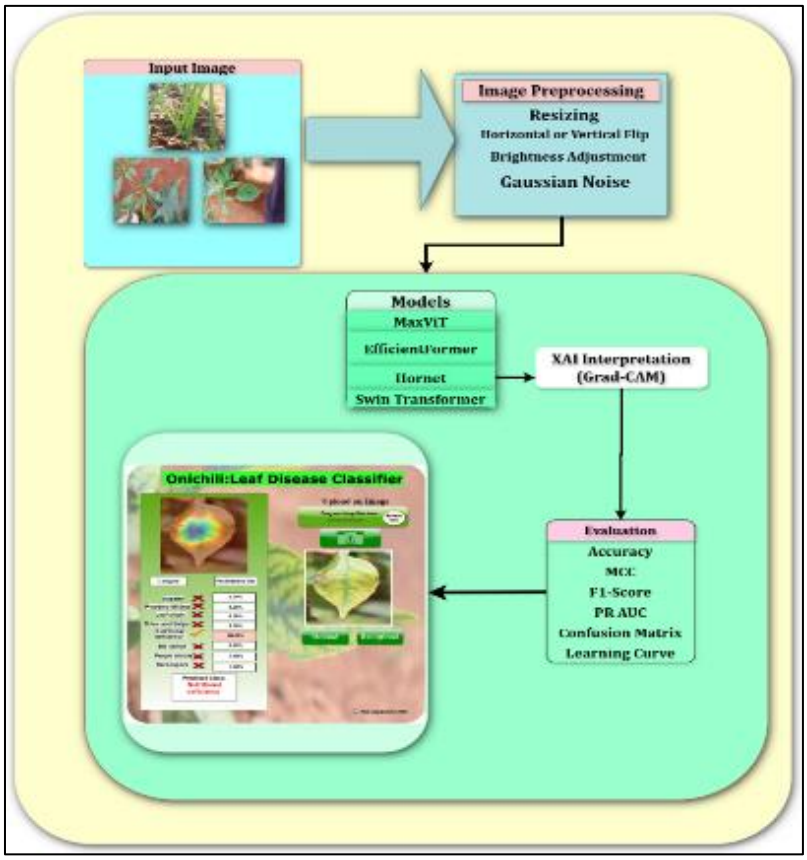

Figure 1 Proposed methodology

3.1 Data Description

The dataset used in this study [15] is the COLD dataset, which consists of high-resolution leaf images of chili and onion plants, specifically collected from the Chilwadigi village in the Koppal district of Karnataka, India. This dataset is designed to support the development of machine learning models for the classification and detection of various leaf

diseases in chili and onion crops. The dataset contains a total of 13,989 images, consisting of 10,987 chili leaf images and 4,502 onion leaf images.

The chili leaf dataset is categorized into five disease classes: healthy leaves (2,198 images), Cercospora leaf spot (2,219 images), mites and thrips (2,507 images), nutritional deficiency (2,032 images), and powdery mildew (2,031 images). The onion leaf dataset is classified into four categories: healthy leaves (1,278 images), Iris yellow spot virus (1,272 images), purple blotch (735 images), and leaf blight (1,217 images). These images were taken during the Kharif season, under varying climatic conditions, and were processed using standard pre-processing techniques such as resizing and augmentation to increase the dataset's diversity. Sample image from each class for both dataset is shown in Figure 2.

The images were captured using a Canon Mark II D digital camera, providing detailed photographs to support accurate disease identification. The dataset includes both raw and augmented images to enhance the robustness of machine learning models and to simulate real-world scenarios in agricultural settings. The images in this dataset cover a wide range of conditions, including subtle inter-class similarities, changes in lighting, and varying background conditions such as different foliage arrangements and light levels. This dataset is made publicly available for research purposes and can be accessed via Mendeley Data repositories, with the dataset links provided for both the chili and onion leaf images. The data can be used for training deep learning models, especially for tasks such as disease classification, disease severity estimation, and leaf segmentation. The rich variety of images in natural environments makes this dataset an essential resource for developing robust disease detection systems tailored to real-world agricultural challenges.

Figure 2 Sample images from both datasets

3.2 Image Preprocessing

The dataset underwent several preprocessing steps to ensure consistency and improve the performance of deep learning models. All images were resized to 224 × 224 pixels to standardize input dimensions. The pixel values were then normalized to the range [0, 1], ensuring better stability and faster convergence during training [16]. To enhance model generalization and address potential overfitting, data augmentation techniques were applied, including random rotations of up to ±30 degrees, horizontal and vertical flips, zooming between 0.9 and 1.1, and brightness adjustments between 0.8 and 1.2. These augmentations simulate real-world conditions, such as lighting variations and leaf orientations. Additionally, Gaussian noise was added to mimic sensor noise. The dataset was split into 80% for training, 15% for validation, and 5% for testing, ensuring balanced and unbiased evaluation. These preprocessing steps helped the model focus on key features while enhancing its robustness across diverse scenarios.

3.3 TL Models

To optimize the performance of the transformer-based models, we employed a consistent set of training parameters across all experiments to ensure a fair comparison. The models were trained for 100 epochs using the Adam optimizer, known for its adaptive learning rate and stable convergence characteristics. The initial learning rate was set at 0.001 and was modified using a learning rate scheduler with a patience of 5 epochs and a decay factor of 0.1. A batch size of 32 was selected to balance memory usage and convergence speed. We applied the categorical cross-entropy loss function, which is suitable for multi-class classification tasks. To prevent overfitting and reduce unnecessary training time, early stopping was implemented with a patience of 10 epochs. Additionally, model checkpointing was utilized to save the best-performing model based on validation accuracy. All experiments were conducted on an NVIDIA GPU with CUDA acceleration, ensuring efficient parallel processing during training. These parameters were empirically selected based on preliminary experiments and align with standard practices in image classification using deep learning.

3.3.1 MaxViT

The MaxViT model shown in Figure 3, combines convolutional layers with multi-scale attention modules to effectively capture both local and global features in high-resolution images. Its foundation lies in the scaled dot-product attention

shown in Equation (1), where Q, K, and V represent the query, key, and value matrices, and d_k is the dimension of the key vectors. The model then employs multi-head attention, as defined in Equation (2), to aggregate diverse feature representations by applying several attention heads in parallel and concatenating their outputs [17]. Additionally, MaxViT introduces local and grid attention mechanisms that merge both fine-grained and global contexts, which can be expressed in Equation (3) by adding the outputs of local ($F_l$ocal) and global ($F_g$lobal) attention modules to the original input X. This fusion enables MaxViT to adapt seamlessly to various disease patterns in leaf images while maintaining computational efficiency [18].

$$\text{Attn} = \text{QKTdkAttn} = \frac{QK^T}{\sqrt{d_k}} \tag{1}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) \cdot W^O \tag{2}$$

$$Y = X + F_{\text{local}}(X) + F_{\text{global}}(X) \tag{3}$$

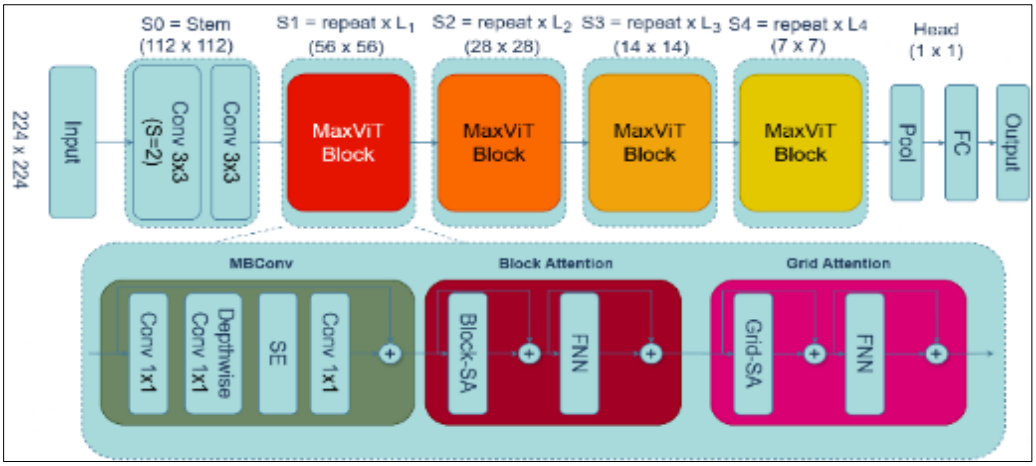

Figure 3 Proposed MaxViT model architecture

3.3.2 Efficient Former

This is a transformer-based model that prioritizes reduced computational overhead while preserving high accuracy. Its core principle involves a hierarchical attention mechanism that selectively processes both local and global features [19]. This is initiated by a lightweight self-attention function, as shown in Equation (4), where Q, K, and V represent the query, key, and value matrices, and d is the feature dimension. The attention scores are aggregated across different heads to form a comprehensive representation of the input. To further enhance efficiency, EfficientFormer introduces a progressive downsampling approach expressed in Equation (5), which systematically reduces spatial resolution at each stage. By doing so, the model focuses on the most relevant details without incurring excessive computational costs [20]. This hierarchical design allows EfficientFormer to excel in detecting small lesions or subtle discolorations in leaf images, making it highly suitable for real-time applications where resources may be limited.

$$\text{LightAttn} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \tag{4}$$

$$\text{Xdown} = \text{DownSample}(X) \tag{5}$$

3.3.3 Hornet

It combines convolutional operations with an attention mechanism to balance efficiency and performance in image classification tasks. It begins by extracting local features through convolution, as indicated in Equation (6), where $\text{Conv}(X)$ denotes the convolution applied to input X. Hornet then refines these features using an attention module described in Equation (7), which takes Q, K, and V from the convolved feature maps and computes attention scores to highlight critical regions. By integrating convolution and attention in a single framework, Hornet captures both small-

scale patterns (such as minor lesions) and broader structural cues (like leaf shape) [21]. This synergy enables Hornet to adapt effectively to diverse disease manifestations in chili and onion leaves, maintaining robust performance under varying environmental conditions. Its balanced computational requirements also make it feasible for real-time agricultural monitoring systems.

$$\text{Xconv} = \text{Conv}(\text{X}) \tag{6}$$

$$\text{HornetAttn} = \text{Softmax}\left(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}}}\right) \cdot \text{V} \tag{7}$$

3.3.4 Swin Transformer

This model employs a hierarchical vision transformer structure, dividing images into non-overlapping windows and applying window-based self-attention at each level. Equation (8) describes the standard attention mechanism within each window, where the input is partitioned, and Q, K, V are computed for that specific region. To broaden the receptive field across layers, Swin implements a shifting strategy in Equation (9), which cyclically shifts windows before the next attention operation. By progressively enlarging the scope of attention, Swin captures both localized features and broader contextual elements, such as overall leaf color or shape [22], [23]. This multi-stage approach balances efficiency and accuracy, enabling the model to handle high-resolution images without excessive computational overhead [24]. Consequently, Swin Transformer is well-suited for detecting various disease symptoms in leaf images, even when they manifest at multiple scales or under complex background conditions.

$$\text{WindowAttn(X)} = \text{Softmax}\left(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}}}\right) \cdot \tag{8}$$

$$\text{Shift(X)} = \text{CyclicShift(X, } \delta\text{)} \tag{9}$$

3.4 Evaluation Metrics

The performance of the multiclass classification model is assessed using four essential evaluation metrics that together offer a comprehensive view of its effectiveness. Accuracy indicates the ratio of correct predictions to the total number of samples, serving as a general performance indicator; however, it may not fully reflect the model's behavior on imbalanced datasets. To address this, the F1-Score, which is the harmonic mean of precision and recall, provides a balanced measure by considering both false positives and false negatives, making it more suitable for datasets with unequal class distribution. The PR-AUC evaluates the trade-off between precision and recall across different thresholds and is particularly helpful in scenarios with class imbalance, as it emphasizes the quality of positive predictions. Lastly, the Matthews Correlation Coefficient offers a single-value summary based on the full confusion matrix, measuring the strength of correlation between predicted and actual labels, and is especially reliable for imbalanced classification problems.

4 Results

4.1 Performance Comparison of Experimental Models

The performance comparison of the deep learning models across chili and onion leaf disease datasets is summarized in Table 1, evaluated using Accuracy, F1-Score, PR-AUC, and MCC. Among all models, MaxViT demonstrates superior performance on both datasets, achieving the highest accuracy 95.25% for onion, 90.48% for chili and MCC 93.97%, 88.86%, reflecting its strong capacity to capture multi-scale disease features. Efficient-Former ranks second, offering a favorable balance between accuracy and computational efficiency, with 92.42% and 88.88% accuracy on onion and chili datasets, respectively. Swin Transformer shows moderate performance but declines on the chili dataset, suggesting sensitivity to intra-class variability. Hornet, while lightweight, reports the lowest scores across all metrics, indicating limited generalization. Overall, MaxViT's consistent top performance highlights its robust feature representation and generalization, making it highly suitable for complex multiclass classification tasks in agricultural image analysis.

Table 1 Performance of experimental models.

Approach	Model	Accuracy	F1	MCC	AUC-PR
Onion Dataset	MaxViT	95.25	94.21	93.97	95.75
	EfficientFormer	92.42	91.31	90.65	93.05
	Swin	90.82	89.45	88.85	91.12
	Hornet	85.40	84.90	83.51	86.20
Chili Dataset	MaxViT	90.48	89.82	88.86	90.86
	EfficientFormer	88.88	87.66	86.98	89.61
	Swin	85.01	84.48	83.70	85.41
	Hornet	81.93	81.78	80.42	82.53

4.2 Performance Validation

The confusion matrices in Figure 4 show the model's strong classification performance across both onion and chili leaf disease datasets. In the onion dataset, most healthy, Iris yellow, purple blotch, and leaf blight samples were correctly identified, with 181, 185, 101, and 174 true predictions respectively. A small number of misclassifications occurred, particularly between purple blotch and other classes, indicating minor visual similarities. In the chili dataset, the model accurately predicted a high number of mites and thrips with 359 correct cases and powdery mildew with 299 correct cases. A few healthy samples were misclassified as Cercospora or powdery mildew, while misclassifications between Cercospora and mites and thrips were limited. Overall, the confusion matrices reflect the model's robust performance and effective feature learning, with only a few errors likely caused by subtle overlaps in disease symptoms.

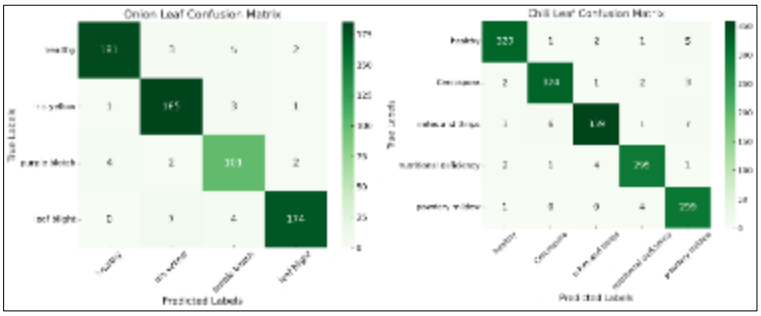

Figure 4 Confusion matrix of the proposed MaxViT model for both datasets

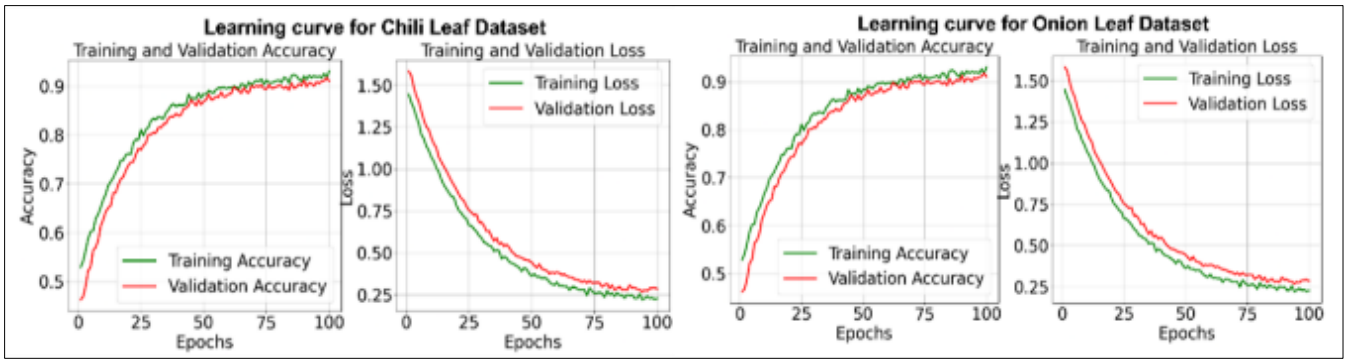

Figure 5 Learning curves of the proposed MaxViT model for both datasets

The learning curves in Figure 5 show the training and validation performance of the proposed model on both the onion and chili leaf datasets over 100 epochs. In the onion dataset, both training and validation accuracy steadily increase and closely converge above 90 percent, indicating good generalization. The corresponding loss curves consistently decrease and stabilize with minimal gaps, suggesting effective learning without overfitting. Similarly, in the chili dataset, training

and validation accuracy rise smoothly with slight fluctuations and converge above 85 percent. The loss curves for chili also show a strong downward trend with both losses aligning well overtime. The overall behavior across both datasets confirms stable convergence, effective optimization, and the model's strong capacity to learn meaningful features while maintaining generalization across classes.

4.3 Model Transparency

Figure 6 compares the Chili Leaf Dataset and Onion Leaf Dataset, showing the original input images (top row) alongside Grad-CAM visualizations (bottom row) for each class. This helps interpret the predictions made by the MaxViT model. Grad-CAM highlights the key areas in the images that influenced the model's decisions, enhancing transparency and understanding of its reasoning.

In the Chili Leaf Dataset, Grad-CAM shows clear activation in specific regions for each category. The Powdery Mildew class highlights white powdery areas, while the Cercospora class focuses on necrotic circular spots. The Nutritional Deficiency class reveals yellow patches and discoloration, and the Mites and Thrips classes indicate insect-affected areas. The Healthy class shows a uniform distribution without stress patterns or lesions. In the Onion Leaf Dataset, Grad-CAM visualizations identify disease-specific markers. The Purple Blotch class highlights a dark central lesion with a purplish surrounding area. The Leaf Blight class shows scattered patterns along the midrib, and the Iris Yellow Spot Virus focuses on yellow striping and discoloration. The Healthy class displays evenly spread heatmaps, indicating no distinct diseases.

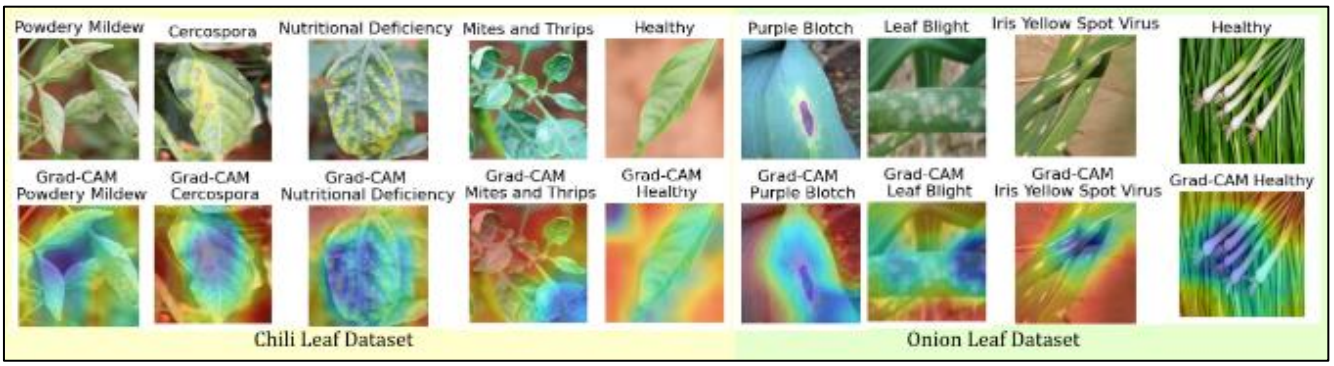

Figure 6 Sample Grad-CAM predictions by MaxViT for each class of both datasets

These visualizations are essential to verify that the model focuses on the right symptomatic areas and allows for human oversight of the classification process. This transparency builds trust, especially in sensitive areas like crop disease monitoring, where it supports timely decision-making and intervention. The Grad-CAM overlays demonstrate that the model can differentiate subtle symptom variations across disease types, indicating effective feature learning and strong class discrimination.

4.4 Web Application

Our Onichili Leaf Disease Classifier web application (Figure 7) is an intelligent diagnostic tool developed using the Django framework, designed to detect and classify common diseases in chili plant leaves. The user-friendly interface enables users to upload leaf images either by dragging and dropping or selecting files via a browser. Upon uploading an image, the application processes it through the proposed model and presents both detailed probability distribution and the final prediction. To enhance interpretability and user trust, the application integrates Grad-CAM visualizations, which generate a heatmap over the leaf image to highlight the specific regions that most influenced the model's decision. This feature offers a transparent and XAI solution for users, particularly beneficial in real-world agricultural diagnosis scenarios. Additional controls such as Reload and Re-upload allow users to test multiple samples with ease. Overall, this application serves as a practical and scalable tool for plant disease management, enabling timely intervention by farmers and agronomists. Django-based architecture ensures flexibility for future enhancements, such as incorporating new disease classes or deploying mobile support for field use.

Figure 7 Onichili web application for plant classification

4.5 State-of-The-Art Comparison

Table 2 presents comparative analysis of previous studies and the proposed MaxViT-based approach in terms of dataset size and classification performance. Most existing works utilized relatively small datasets, ranging from 400 to 3200 samples, and achieved accuracy scores between 86.2% and 95.0%. Notably, the MobileNetV1 and EfficientNetB4-based models demonstrated competitive performance with 95.0% and 92.5% accuracy, respectively. In contrast, the proposed MaxViT model was evaluated on a significantly larger dataset containing 13,989 samples, covering both chili and onion leaf diseases. It achieved superior results, with 95.75% accuracy on the onion dataset and 90.86% on the chili dataset. These results not only validate the robustness of MaxViT in handling complex multiclass classification tasks but also demonstrate its effectiveness on large-scale agricultural datasets. The improved performance underscores the model's strong generalization capabilities and its potential applicability in real-world plant disease detection scenarios.

Table 2 Performance comparison with previous studies

Model	Data sample size	Result (%)
EfficientNetB4[4]	2560	92.5
MobileNetV1[5]	3200	95
UAV-NIR [6]	400	89.7
ML-based [7]	450	86.2
SPA-BP [8]	480	92.3
(Our) MaxViT	13,989	95.75, 90.86

5 Discussion

This study demonstrates the effectiveness of transformer-based deep learning models for automated detection of chili and onion leaf diseases. Among all evaluated models, MaxViT achieved the highest performance, with 95.75% accuracy on the onion dataset and 90.86% on the chili dataset. Its hybrid architecture, combining convolutional and self-attention mechanisms, enabled it to capture both fine-grained and global disease patterns, leading to better generalization across complex leaf textures. EfficientFormer also showed strong results, offering a favorable balance between accuracy and computational efficiency. Carefully designed preprocessing steps such as resizing, normalization, and Gaussian noise injection improved feature consistency and model robustness under varied image conditions. The integration of the trained model into a real-time web application further enhances its practical impact, providing instant predictions for agricultural stakeholders. Importantly, the use of Grad-CAM-based XAI improves interpretability by visually highlighting the disease-affected regions, fostering trust in model outputs. Despite its strengths, the system faces limitations including reduced performance on visually overlapping chili diseases and the high computational cost of

transformer models. Future work should focus on dataset expansion, model optimization for edge deployment, and richer interpretability techniques for real-world agricultural adoption.

6 Conclusion

This study presents a transformer-based deep learning framework for detecting multiple leaf diseases in chili and onion plants, using high-performing models such as MaxViT, Swin, Hornet, and EfficientFormer. Through careful preprocessing and model selection, we achieved strong classification results across both datasets. To support real-world applications, we developed a lightweight and accessible web application that enables real-time disease prediction directly from user-uploaded images. The integration of XAI through Grad-CAM further enhances the system by visually identifying key disease regions, promoting transparency and trust in model decisions. This work bridges the gap between research and practical deployment, offering a scalable solution for improving plant health monitoring in agricultural settings. Nonetheless, the study is limited by the scope of disease types and the computational demands of transformer architectures. Future efforts will focus on expanding the dataset, optimizing models for mobile deployment, and improving interpretability. These steps aim to strengthen the usability and impact of AI in precision agriculture.

Compliance with ethical standards

Disclosure of conflict of interest

There is not conflict of interests.

References

[1] J. D. Teshika et al., "Traditional and modern uses of onion bulb (Allium cepa L.): a systematic review," Crit Rev Food Sci Nutr, vol. 59, no. Sup 1, pp. S39–S70, 2019, doi: 10.1080/10408398.2018.1499074.

[2] S. Nazeer, T. T. R. Afzal, Sana, M. Saeed, S. Sharif, and M. Zia-Ul-Haq, "Chili Pepper," Essentials of Medicinal and Aromatic Crops, pp. 855–885, 2023, doi: 10.1007/978-3-031-35403-8_33.

[3] R. Haque et al., "Advancements in Jute Leaf Disease Detection: A Comprehensive Study Utilizing Machine Learning and Deep Learning Techniques," in PEEIACON 2024 - International Conference on Power, Electrical, Electronics and Industrial Applications, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 248–253. doi: 10.1109/PEEIACON63629.2024.10800378.

[4] L. Zhang and X. Wu, "The Lightweight Deep Learning Model in Sunflower Disease Identification: A Comparative Study," Applied Sciences (Switzerland), vol. 15, no. 4, Feb. 2025, doi: 10.3390/app15042104.

[5] Md. R. Ahmed et al., "Towards Automated Detection of Tomato Leaf Diseases," 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), pp. 387–392, May 2024, doi: 10.1109/ICEEICT62016.2024.10534559.

[6] R. Haque et al., "Advancements in Jute Leaf Disease Detection: A Comprehensive Study Utilizing Machine Learning and Deep Learning Techniques," 2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON), pp. 248–253, Sep. 2024, doi: 10.1109/PEEIACON63629.2024.10800378.

[7] K. Kavitha and S. Naveena, "Deep Learning Framework for Identification of Leaf Diseases in Native Plants of Tamil Nadu Geographical Region," in 2023 International Conference on Computer Communication and Informatics, ICCCI 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCCI56745.2023.10128593.

[8] M. S. Rahman et al., "Effective Disease Recognition in Cucumbers: A Web-Based Application Using Transfer Learning Models," 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2024 - Proceedings, pp. 59–64, 2024, doi: 10.1109/RAAICON64172.2024.10928353.

[9] A. Al-Sakib, F. Islam, R. Haque, M. B. Islam, A. Siddiqua, and M. M. Rahman, "Classroom Activity Classification with Deep Learning," 2nd International Conference on Integrated Circuits and Communication Systems, ICICACS 2024, 2024, doi: 10.1109/ICICACS60521.2024.10498187.

[10] V. K. Pratap and N. Suresh Kumar, "High-precision multiclass classification of chili leaf disease through customized EffecientNetB4 from chili leaf images," Smart Agricultural Technology, vol. 5, Oct. 2023, doi: 10.1016/j.atech.2023.100295.

[11] R. Muslim, Z. Zaeniah, A. Akbar, B. Imran, and Z. Zaenudin, "Disease Detection of Rice and Chili Based on Image Classification Using Convolutional Neural Network Android-Based," Jurnal Pilar Nusa Mandiri, vol. 19, no. 2, pp. 85–96, Sep. 2023, doi: 10.33480/pilar.v19i2.4669.

[12] M. R. McDonald, C. S. Tayviah, and B. D. Gossen, "Human vs. Machine, the Eyes Have It. Assessment of Stemphylium Leaf Blight on Onion Using Aerial Photographs from an NIR Camera," Remote Sens (Basel), vol. 14, no. 2, Jan. 2022, doi: 10.3390/rs14020293.

[13] M. Amondkar, V. Karad, and S. Bhoite, "Machine Learning Approach for Onion Leaf Disease Detection: A Case Study in Maharashtra, India Sachin Bhimrao Bhoite Machine Learning Approach for Onion Leaf Disease Detection: A Case Study in Maharashtra, India." [Online]. Available: https://www.researchgate.net/publication/383063286

[14] Y. Shao et al., "Detection and Analysis of Chili Pepper Root Rot by Hyperspectral Imaging Technology," Agronomy, vol. 14, no. 1, Jan. 2024, doi: 10.3390/agronomy14010226.

[15] M. P. Aishwarya and A. P. Reddy, "Dataset of chilli and onion plant leaf images for classification and detection," Data Brief, vol. 54, Jun. 2024, doi: 10.1016/j.dib.2024.110524.

[16] M. D. Hosen et al., "Parasitology Unveiled: Revolutionizing Microorganism Classification Through Deep Learning," 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), pp. 1163–1168, May 2024, doi: 10.1109/ICEEICT62016.2024.10534322.

[17] M. Sarıateş and E. Özbay, "A Classifier Model Using Fine-Tuned Convolutional Neural Network and Transfer Learning Approaches for Prostate Cancer Detection," Applied Sciences 2025, Vol. 15, Page 225, vol. 15, no. 1, p. 225, Dec. 2024, doi: 10.3390/APP15010225.

[18] A. Sriwastawa and J. A. Arul Jothi, "Vision transformer and its variants for image classification in digital breast cancer histopathology: a comparative study," Multimed Tools Appl, vol. 83, no. 13, pp. 39731–39753, Apr. 2024, doi: 10.1007/S11042-023-16954-X/METRICS.

[19] E. B. Hamdi and Hidayaturrahman, "Ensemble of pre-trained vision transformer models in plant disease classification, an efficient approach," Procedia Comput Sci, vol. 245, no. C, pp. 565–573, Jan. 2024, doi: 10.1016/J.PROCS.2024.10.283.

[20] J. Sa, J. Ryu, and H. Kim, "ECTFormer: An efficient Conv-Transformer model design for image recognition," Pattern Recognit, vol. 159, p. 111092, Mar. 2025, doi: 10.1016/J.PATCOG.2024.111092.

[21] S. Sun, Z. Gao, C. Huang, and H. Yu, "GloVe-FRCNN: Comprehensive network algorithm for Vespa mandarinia image-text extraction and classification," 2021 IEEE 3rd International Conference on Communications, Information System and Computer Engineering, CISCE 2021, pp. 349–355, May 2021, doi: 10.1109/CISCE52179.2021.9445902.

[22] S. Zhang, K. Zhao, Y. Huo, M. Yao, L. Xue, and H. Wang, "Mushroom image classification and recognition based on improved ConvNeXt V2," J Food Sci, vol. 90, no. 3, p. e70133, Mar. 2025, doi: 10.1111/1750-3841.70133.

[23] J. Zhang, H. Zhou, K. Liu, and Y. Xu, "ED-Swin Transformer: A Cassava Disease Classification Model Integrated with UAV Images," Sensors 2025, Vol. 25, Page 2432, vol. 25, no. 8, p. 2432, Apr. 2025, doi: 10.3390/S25082432.

[24] F. Huo, H. Li, H. Dong, and W. Ren, "An improved Swin transformer for sandstone micro-image classification," Geoenergy Science and Engineering, vol. 247, p. 213680, Apr. 2025, doi: 10.1016/J.GEOEN.2025.213680.