

AgriVision: A Vision Transformer-Based System for Plant Disease Detection and Advisory

Ansa Scaria¹, Nandana Ashok², Navya S Nair³,
Sebin Binoy⁴, Ojus Thomas Lee⁵, Shandry K K⁶

Department of CSE, College of Engineering Kidangoor, Kerala, India

1 ansascaria@gmail.com, 2 nandhanaashok2004@gmail.com, 3 navyasnair05@gmail.com, 4
sebinbinoysebin@gmail.com, 5 ojustl@ce-kgr.org, 6 shandrykk@ce-kgr.org

Abstract - Plant diseases pose a significant threat to global agricultural productivity, leading to substantial economic losses and food security challenges. Traditional disease detection methods rely heavily on manual inspection by experts, which is timeconsuming, labor-intensive, and often impractical for large-scale farming. Recent advancements in deep learning have enabled automated disease detection; however, most existing systems are limited to standalone classification models without real-world integration.

This paper presents AgriVision, a Vision Transformer (ViT)-based plant disease detection and advisory system. The proposed approach leverages the self-attention mechanism of transformers to capture global contextual features from plant leaf images, improving classification performance compared to conventional Convolutional Neural Networks (CNNs). In addition to disease prediction, the system integrates an officer validation module to enhance reliability and a marketplace module to provide farmers with access to recommended agricultural products. The model is trained and evaluated on multiple plant disease datasets using standard performance metrics such as accuracy, precision, recall, and F1-score. Experimental results show that the proposed model achieves a test accuracy of 94.21%, with a precision of 94.48%, recall of 94.21%, and F1-score of 94.01%, outperforming the baseline CNN model. The overall system provides an end-to-end solution that bridges the gap between disease detection, diagnosis, and treatment, contributing to the development of scalable smart agriculture systems.

Key Words Plant disease detection, Vision Transformer, Deep Learning, Smart Agriculture, Computer Vision

1. INTRODUCTION

Agriculture plays a crucial role in sustaining global food security and economic development. However, plant diseases remain a major challenge, causing significant crop losses and adversely affecting farmer livelihoods worldwide. Early and accurate disease detection is essential to prevent the spread of infections and enable timely intervention. Traditional detection methods rely on manual inspection by agricultural experts, which is time-consuming, labor-intensive, and impractical for large-scale monitoring.

With the rapid advancement of artificial intelligence, deep learning techniques have emerged as effective solutions for automated plant disease detection. Convolutional Neural Networks (CNNs) have been widely used for image-based classification tasks and have demonstrated promising results in identifying plant diseases [1]. However, CNNs are inherently

limited in capturing long-range dependencies and global contextual information due to their localized receptive fields.

Recently, Vision Transformers (ViTs) have gained significant attention in computer vision tasks due to their ability to model global relationships using self-attention mechanisms [2], [4]. Unlike CNNs, ViTs process images as sequences of patches, enabling them to capture both local and global features effectively. This makes them particularly suitable for complex image classification problems such as plant disease detection.

Despite these advancements, there remains a gap between model-level predictions and real-world agricultural applications. Most existing solutions focus primarily on disease classification and do not provide integrated support for validation, treatment guidance, and accessibility of agricultural resources.

To address these limitations, this paper proposes AgriVision, a Vision Transformer-based plant disease detection and advisory system. The proposed system integrates multiple components, including an automated disease detection module, an officer validation mechanism to improve prediction reliability, and a marketplace module to connect farmers with relevant agricultural products. This end-to-end approach enhances both detection performance and practical usability in real-world scenarios.

The main contributions of this work are as follows:

- Development of a Vision Transformer-based model for accurate plant disease classification.
- Design of an integrated framework combining disease detection, validation, and advisory functionalities.
- Implementation of a web-based platform for real-time disease diagnosis and treatment support.

2. RELATED WORK

Plant disease detection has been widely explored using deep learning techniques, particularly Convolutional Neural Networks (CNNs). Early studies [1] utilized pre-trained architectures such as VGG, ResNet, and Inception for feature extraction and classification, achieving high accuracy through transfer learning [1]. These approaches demonstrated the effectiveness of deep learning in plant disease identification; however, they primarily relied on local feature extraction.

To improve efficiency and performance, EfficientNet-based models were introduced, which optimize network scaling in terms of depth, width, and resolution [3]. While these models achieve better accuracy with fewer parameters, they still face limitations in capturing global contextual relationships within images.

Recently, Vision Transformers (ViTs) have emerged as a powerful alternative for image classification tasks. ViTs process images as sequences of patches and utilize self-attention mechanisms to model long-range dependencies [2], [4]. This enables them to capture both local and global features more effectively compared to traditional CNN-based approaches. Several studies have demonstrated that ViT-based models outperform CNNs in plant disease detection tasks, particularly in complex scenarios.

In addition, hybrid architectures that combine CNNs and Vision Transformers have been proposed to leverage the advantages of both methods [5]. These approaches improve feature representation by integrating local feature extraction with global attention mechanisms.

Despite these advancements, most existing works focus primarily on improving classification accuracy and do not address the practical challenges of real-world agricultural applications. In particular, they lack integration with systems that support expert validation, treatment recommendation, and accessibility to agricultural resources.

The proposed AgriVision system addresses these limitations by combining a Vision Transformer-based detection model with an integrated platform that includes an officer validation module and a marketplace component. This approach provides not only accurate disease detection but also practical support for real-time agricultural decision-making, addressing the limitations of existing methods.

3. PROPOSED SYSTEM

A. System Overview

The proposed system, AgriVision, is an intelligent plant disease detection and advisory platform designed to assist farmers in early disease identification and treatment. The system integrates deep learning techniques with a web-based application to provide an end-to-end solution for crop disease management.

Unlike conventional approaches that focus solely on classification, the proposed system combines automated disease detection with expert validation and advisory support. This integration enhances both the accuracy and practical usability of the system in real-world agricultural scenarios.

B. System Architecture

Fig. 1 illustrates the overall architecture and interaction between different modules of the proposed system.

- Data Acquisition and Preprocessing Module
- Disease Detection Module (Vision Transformer)
- Officer Validation Module
- Marketplace Module

The system architecture is designed to ensure smooth data flow from image input to final recommendation output.

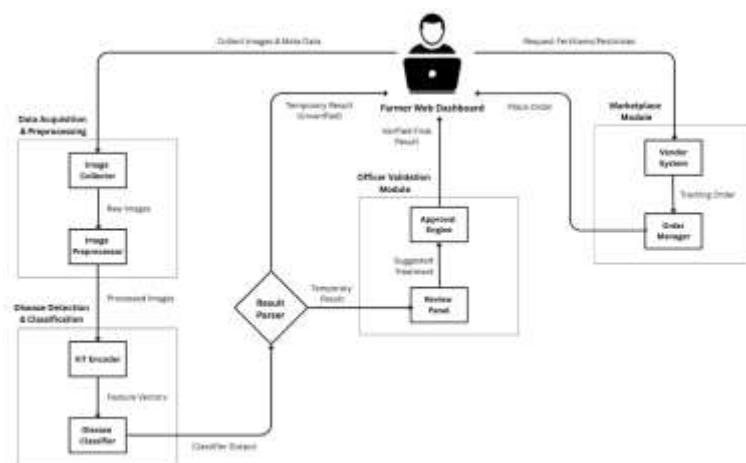


Fig. 1: System Architecture of the Proposed AgriVision System.

C. Modules Description

1. Data Acquisition and Preprocessing Module: This module handles the collection and preparation of input images from users.

- Users upload plant leaf images through the web interface.
- The system validates the uploaded image format and quality.
- Images are resized to a fixed resolution of 224×224 pixels.
- Pixel values are normalized to ensure consistent input distribution.
- Data augmentation techniques such as rotation, flipping, and scaling are applied to improve model generalization.

2. Disease Detection Module: This module performs automated plant disease classification using a Vision Transformer (ViT) model.

- The input image is divided into fixed-size patches.
- Each patch is converted into an embedding vector using linear projection.
- Positional encoding is added to retain spatial information.
- The sequence of embeddings is passed through transformer encoder layers.
- Multi-head self-attention is used to capture global dependencies across image regions.
- The final representation is processed through a classification head to predict the disease class.
- The system outputs the predicted disease along with confidence scores.

Fig. 2 shows a sample detection result of the Disease Detection Module. A diseased plant leaf image is provided as input, and the system outputs the predicted disease class along with a confidence score.



Fig. 2: Sample Output of the Disease Detection Module

3. Officer Validation Module: This module ensures the reliability of predictions through expert verification.

- The predicted result is forwarded to an agricultural officer.
- The officer reviews the input image and predicted class.
- The prediction is either approved or corrected based on expert knowledge.
- The validated result is stored and sent back to the user.

4. Marketplace Module: This module provides users with access to recommended agricultural products.

- The system suggests fertilizers and pesticides based on the detected disease.
- Users can view product details and recommendations.
- The platform enables users to connect with vendors for purchasing products.

D. System Workflow

The workflow of the proposed system is as follows:

- 1) The user uploads an image of a plant leaf through the interface.
- 2) The image is preprocessed and passed to the Vision Transformer model.
- 3) The model generates an initial disease prediction, which is immediately displayed to the user.
- 4) The prediction is then forwarded to an agricultural officer for validation.
- 5) After verification, the validated result along with treatment recommendations is provided to the user.
- 6) The user can access the marketplace to purchase recommended agricultural products.

This dual-stage prediction mechanism ensures both rapid response and improved reliability of the system.

4. METHODOLOGY

A. Overview

The proposed system utilizes a Vision Transformer (ViT) model for plant disease detection. Unlike traditional Convolutional Neural Networks (CNNs), which rely on convolution operations, the Vision Transformer processes

images as sequences of patches and captures global relationships using self-attention mechanisms.

B. Data Preprocessing

Input images are resized to 224×224 pixels and normalized to match the input requirements of the Vision Transformer model. The datasets used in this study include both naturally diverse images and pre-augmented samples. No additional data augmentation is applied during training, as the existing dataset diversity is sufficient for effective model learning.

C. Model Architecture

1) Baseline Model- CNN: The baseline Convolutional Neural Network (CNN) model is designed to perform feature extraction and classification in a hierarchical manner. It consists of two convolutional blocks, each aimed at capturing increasingly complex spatial features from the input image.

Each convolutional block contains two Conv2D layers with 3×3 kernels and Rectified Linear Unit (ReLU) activation functions, which introduce non-linearity and enable the model to learn complex patterns. The first block employs 32 filters to extract low-level features such as edges and textures, while the second block uses 64 filters to capture higher-level and more abstract representations of plant disease characteristics.

Following each convolutional block, a Max Pooling layer is applied to reduce the spatial dimensions of the feature maps, thereby decreasing computational complexity and helping to retain the most prominent features. After the second pooling operation, the resulting feature map has a dimension of 56×56×64.

This feature map is then flattened into a one-dimensional vector and passed through a fully connected (dense) layer with 128 neurons, which learns a compact representation for classification. Finally, a Softmax output layer is used to

generate the probability distribution over the target disease classes, enabling multi-class classification.

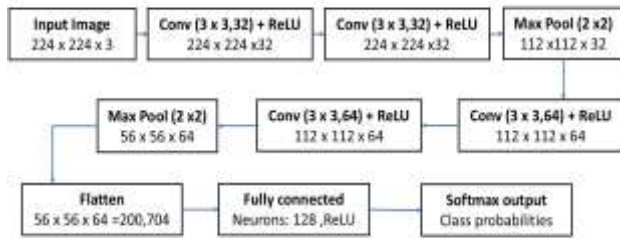


Fig. 3: CNN architecture showing feature extraction and classification stages

2) Proposed ViT Model: The proposed AgriVision system employs a Vision Transformer (ViT) as the core classification model. Unlike conventional convolutional architectures, the ViT processes the input image as a sequence of fixed-size patches and models global relationships using a self-attention mechanism.

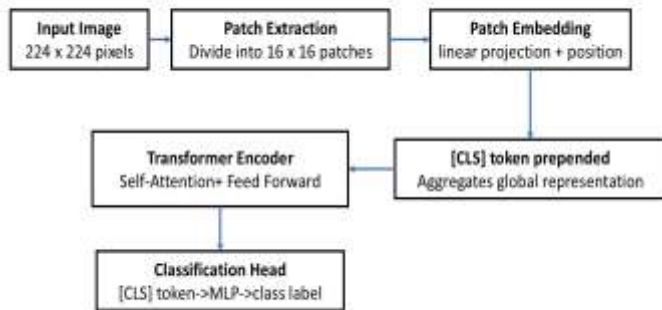


Fig. 4: Architecture of the Proposed Vision Transformer Model

The input image of size 224×224 pixels is divided into non-overlapping patches of size 16×16 , resulting in a total of 196 patches. Each patch is flattened and projected into a 768-dimensional embedding vector using a linear projection layer. To retain spatial information, positional embeddings are added to the patch embeddings. Additionally, a learnable classification token ([CLS]) is prepended to the sequence, forming an input sequence of 197 tokens.

The sequence is passed through a stack of 12 Transformer encoder layers. Each encoder layer consists of a Multi-Head Self-Attention (MHSA) mechanism followed by a Feed Forward Network (FFN), along with residual connections and layer normalization. The self-attention mechanism enables the model to capture long-range dependencies by allowing each patch to attend to all other patches in the image.

After processing through the encoder layers, the final representation of the [CLS] token is extracted and passed through a fully connected layer, followed by a Softmax activation function to produce the probability distribution over the disease classes.

5. IMPLEMENTATION

A. Dataset

The proposed system is trained using three publicly available plant disease datasets, including PlantDoc, Chilli Plant Diseases Dataset, and Coconut Diseases and Pest Infestations Dataset.

The combined dataset consists of approximately 14,446 images across multiple disease categories. The use of multiple datasets improves diversity in terms of plant species, disease types, and real-world variations, enhancing the robustness of the model.

TABLE I: Dataset Summary

Dataset	No. of Classes	Description
PlantDoc [11]	27	Real field images
Chilli Dataset [12]	8	Healthy and diseased chili leaves
Coconut Dataset [13]	6	Pest and disease samples

B. Preprocessing

All images are resized to 224×224 pixels and normalized to ensure consistent input distribution. The Chilli dataset used in this work is pre-augmented by the dataset provider, including variations such as rotation, flipping, and scaling. Therefore, no additional augmentation techniques are applied during training.

This approach simplifies preprocessing while maintaining sufficient variability for model generalization.

C. Training Setup

The combined dataset is divided into training and testing sets using an 80:20 ratio. A portion of the training data is further used for validation during model training to monitor performance and prevent overfitting.

The Vision Transformer model is implemented using the HuggingFace Transformers library and trained using the PyTorch framework. The training process involves iterative optimization of model parameters using backpropagation and gradient descent.

TABLE II: Training Configuration

Parameter	Value
Model	Vision Transformer (ViT)
Image Size	224×224
Batch Size	16
Epochs	10
Learning Rate	0.00005
Optimizer	AdamW
Framework	PyTorch, HuggingFace
Training Platform	Kaggle

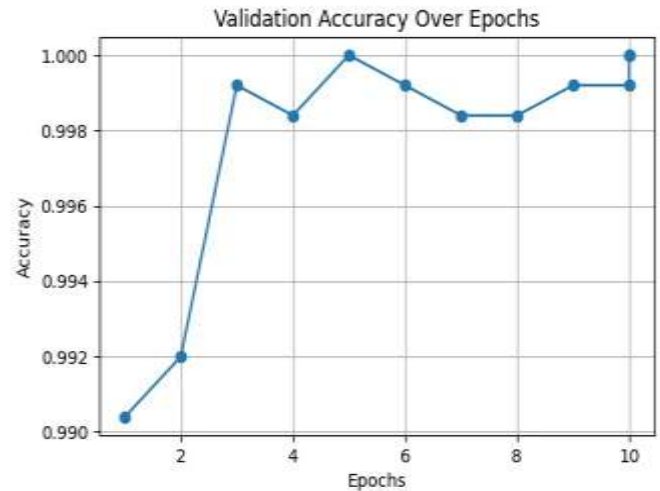


Fig. 5: Validation Accuracy of the Proposed Model

D. Training Environment

The model is trained on the Kaggle platform using cloudbased GPU resources, enabling efficient training of the deep learning model. Supporting libraries such as NumPy, scikitlearn, and torchvision are used for data processing and evaluation.

The trained model is integrated into a web-based application to provide real-time plant disease detection and advisory services. This integration ensures practical usability of the system beyond model-level predictions.

6. . RESULTS AND DISCUSSION

A. Performance Evaluation

The performance of the proposed Vision Transformer model is evaluated using standard classification metrics including accuracy, precision, recall, and F1-score. Accuracy represents the overall correctness of predictions, while precision and recall measure the quality and completeness of the classification. The F1-score provides a balanced evaluation by combining both precision and recall.

TABLE III: Performance of the Proposed Vision Transformer Model

Metric	Value
Accuracy	94.21%
Precision	94.48%
Recall	94.21%
F1-Score	94.01%

The model achieves strong and consistent performance across all evaluation metrics, demonstrating its effectiveness in multi-class plant disease classification involving 41 classes.

Fig. 5 illustrates the validation accuracy over training epochs

The accuracy curve shows a steady increase during training followed by convergence, indicating that the model learns meaningful feature representations without significant overfitting. The consistent performance suggests good generalization capability on unseen data.

B. Comparison with CNN Models

Convolutional Neural Networks (CNNs), such as ResNet50, are widely used for plant disease classification due to their effectiveness in extracting local features. However, CNNs are inherently limited in capturing global contextual relationships within images.

In contrast, the Vision Transformer leverages a self-attention mechanism that enables the model to process all image patches simultaneously and capture long-range dependencies. This capability allows the model to better identify disease patterns distributed across different regions of the leaf, making it more suitable for complex multi-class classification tasks.

The superior performance of the proposed Vision Transformer model highlights its effectiveness in handling finegrained disease variations compared to conventional CNNbased approaches.

C. Analysis

Fig. 6 presents the confusion matrix restricted to misclassified disease classes. Since the model achieves a high overall accuracy, most classes are correctly predicted and are excluded for clarity.

The observed misclassifications are primarily concentrated among visually similar disease categories within the same plant species. For example, confusion occurs between *Corn Gray Leaf Spot* and *Corn Leaf Blight*, as both diseases exhibit similar lesion patterns. Similarly, *Potato Early Blight* and *Potato Late Blight* show overlapping visual characteristics, particularly in early stages.

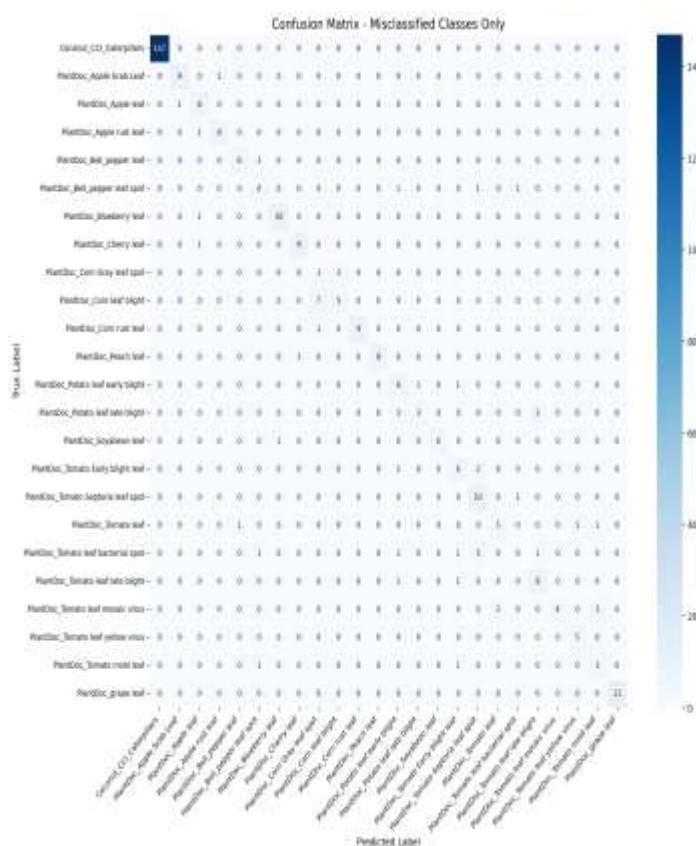


Fig. 6: Confusion Matrix of Misclassified Disease Classes. Only classes with prediction errors are shown to highlight challenging

Among tomato diseases, minor confusion is observed between *Tomato Mosaic Virus*, *Tomato Mold Leaf*, and *Tomato Septoria Leaf Spot*, which share similar discoloration and spotting patterns.

These misclassifications are expected in fine-grained classification tasks and highlight the inherent difficulty of distinguishing visually similar plant diseases. The integration of the officer validation module in the AgriVision system helps mitigate this limitation by enabling expert verification, thereby improving the reliability of real-world predictions.

7.CONCLUSION

This paper presented AgriVision, a Vision Transformerbased intelligent system for plant disease detection and advisory. The proposed system integrates deep learning techniques with a real-time web platform to provide accurate disease classification and practical support for farmers. The use of Vision Transformers enables effective modeling of global contextual features, resulting in improved performance compared to traditional CNN-based approaches. In addition to accurate prediction, the system incorporates an officer validation module to enhance reliability and a marketplace module to provide farmers with direct access to recommended agricultural products. The experimental results demonstrate that the proposed system achieves high accuracy and is capable of identifying plant diseases efficiently. The integration of

detection, validation, and advisory functionalities makes the system suitable for realworld agricultural applications. The system performance may vary depending on dataset diversity and real-world environmental conditions such as lighting, background variations, and image quality. Future work will focus on expanding the system by incorporating larger and more diverse datasets, improving model efficiency for mobile deployment, and integrating additional features such as weather-based disease prediction and multilingual support for farmers.

REFERENCES

1. P. Bansal, R. Kumar, and S. Kumar, "Disease Detection in Apple Leaves Using Deep Convolutional Neural Network," *Agriculture*, vol. 11, no. 7, pp. 617, 2021.
2. R. A. Boukabouya, A. Moussaoui, and M. Berrimi, "Vision Transformer Based Models for Plant Disease Detection and Diagnosis," in *Proc. International Symposium on Informatics and its Applications (ISIA)*, 2022.
3. G. Li, Y. Wang, Q. Zhao, P. Yuan, and B. Chang, "PMVT: A Lightweight Vision Transformer for Plant Disease Identification on Mobile Devices," *Frontiers in Plant Science*, vol. 14, 2023.
4. D. Brown and M. De Silva, "Plant Disease Detection on Multispectral Images using Vision Transformers," *Rhodes University, South Africa*, 2023.
5. H. H. Alshammari and H. Alkhiri, "Optimized Recurrent Neural Network Mechanism for Olive Leaf Disease Diagnosis Based on Wavelet Transform," *Alexandria Engineering Journal*, vol. 78, pp. 149–161, 2023.
6. S. A. Alshammari, "TLDViT: A Vision Transformer Model for Tomato Leaf Disease Classification," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 12, 2024.
7. U. Barman, P. Sarma, M. Rahman, V. Deka, S. Lahkar, and M. J. Saikia, "ViT-SmartAgri: Vision Transformer and Smartphone-Based Plant Disease Detection for Smart Agriculture," *Agronomy*, 2024.
8. H. Kukadiya, D. Meva, N. Arora, and S. Srivastava, "Effective Groundnut Crop Management by Early Prediction of Leaf Diseases through Convolutional Neural Networks," *International Research Journal of Multidisciplinary Technovation*, vol. 6, no. 1, pp. 17–31, 2024.
9. V. V. Gahukar, M. Ramteke, and T. Yengantiwar, "AiKrishi Crop Disease Detection Using Vision Transformer," *International Journal for Research in Applied Science & Engineering Technology*, vol. 13, no. 4, 2025.
10. M. Ali, S. Mouatassim, M. El Haji, and J. Benhra, "Plant Disease Detection using Vision Transformers," *International Journal of Electrical and Computer Engineering*, 2025.
11. PlantDoc Dataset. [Online]. Available: <https://www.kaggle.com/datasets/abdulhasibuddin/plant-doc-dataset>
12. Chilli Plant Diseases Dataset. [Online]. Available: <https://www.kaggle.com/datasets/ravindubandara3002/chilli-plant-diseases-dataset>
13. Coconut Diseases Dataset. [Online]. Available: <https://www.kaggle.com/datasets/samitha96/coconutdiseases>