Discover Artificial Intelligence

Research

Enhancing sugarcane leaf disease classification using vision transformers over CNNs

Saritha Miryala[1] · **Krupa Rasane**[2]

Received: 1 February 2025 / Accepted: 23 May 2025
Published online: 03 June 2025

Abstract
Sugarcane is a globally significant crop facing threats from leaf diseases that impact its productivity. Traditional detection methods are often inefficient and time-consuming. This study explores the use of Vision Transformers (ViT) for classifying sugarcane leaf diseases and compares their performance with traditional CNNs. A dataset of 19,926 images across six classes was used to fine-tune both ViT and CNN models. The optimized ViT model achieved a test accuracy of 96.53%, outperforming the CNN models (ResNet50 and VGG16) with accuracies of 91.92% and 92.30%, respectively. These findings demonstrate the superior performance of ViTs over CNNs in early disease detection for sustainable crop management. Future work will focus on expanding the dataset and optimizing model parameters for further improvements in disease classification accuracy.

Keywords Sugarcane leaf disease detection · CNN · Deep learning · ViT · Image classification

1 Introduction

Sugarcane (Saccharum officinarum) is a crucial cash crop in the global economy, serving as a primary source of sugar and biofuels. However, leaf diseases significantly impact sugarcane productivity and quality, posing major challenges for farmers and the agricultural industry. Early detection of these diseases is essential for minimizing their detrimental effects and ensuring sustainable farming practices [1]. Traditional methods for diagnosing plant diseases, such as manual inspections and laboratory analyses, are often slow, labor-intensive, and prone to human error [1, 2]. These limitations underscore the need for automated, accurate, and efficient disease detection techniques.

Advancements in digital technology and artificial intelligence have facilitated the development of innovative tools to address agricultural challenges. Computer vision, empowered by image processing and deep learning methodologies, offers a promising approach to non-invasive and scalable disease identification [3, 4]. CNN have demonstrated strong performance in various agricultural applications, including crop disease classification, achieving accuracy rates of 90–95% in several studies [4, 5]. CNNs effectively learn hierarchical features from images, making them widely employed for plant disease detection. However, they often require large amounts of annotated data and struggle with modeling long-range dependencies within images [6]. Additionally, CNNs are sensitive to variations in scale, orientation, and background clutter, which can reduce classification accuracy in real-world agricultural settings [7, 8].

Deep learning has been widely adopted for plant disease detection across various crops. For instance, Kaur and Gupta [2] utilized lightning-fast CNNs for automating potato leaf disease detection, achieving significant improvements in precision.

✉ Saritha Miryala, saritha.jce@gmail.com | [1]Department of Artificial Intelligence and Data Science, S.G. Balekundri Institute of Technology, Belagavi, India. [2]Department of Electronics and Communication Engineering, Jain College of Engineering, Belagavi, India.

Similarly, Shruthi et al. [4] reviewed machine learning classification techniques for plant disease detection, highlighting the effectiveness of deep learning models in agriculture. Saleem et al. [5] explored deep learning methods for plant disease detection and classification, demonstrating the potential of these techniques to handle complex datasets and improve accuracy. Demilie [6] conducted a comparative study on plant disease detection techniques, emphasizing the superior performance of deep learning approaches over traditional methods.

ViTs have recently emerged as a powerful alternative to CNNs for image recognition tasks [9]. ViTs leverage the self-attention mechanism from Transformer architectures, originally developed for natural language processing [10], to model global relationships within image data. By segmenting images into sequences of patches and processing them similarly to words in a sentence, ViTs can capture long-range dependencies and intricate patterns more effectively than CNNs [11]. This capability is especially beneficial in plant disease detection, where symptoms may present subtle visual differences across different regions of the leaf [12].

Recent studies have begun to explore the application of ViTs in agricultural contexts. Borhani et al. [11] proposed a deep learning-based approach for automated plant disease classification using Vision Transformers, achieving promising results. Boukabouya et al. [12] developed ViT-based models for plant disease detection and diagnosis, highlighting the potential of these architectures in agriculture. Barman and Sarma [13] introduced ViT-SmartAgri, a smartphone-based plant disease detection system utilizing Vision Transformers, further demonstrating the practicality of ViTs in agricultural applications.

Recent advancements in deep learning have significantly impacted agriculture, extending beyond sugarcane to a variety of crops. In a study by Kunduracıoğlu and Paçal [14], Vision Transformers (ViTs) demonstrated superior performance in the classification of sugarcane leaf diseases, surpassing traditional CNN-based methods. Similarly, the application of deep learning models to the classification of tomato diseases using ResNet architectures has shown impressive results, with Res2 Next50 achieving a 99.85% accuracy rate [15]. Additionally, advancements in deep learning for grape leaf disease classification have further exemplified the potential of these models in agriculture, with pre-trained CNNs and ViTs achieving 100% accuracy on certain datasets [16]. These studies collectively highlight the growing importance of deep learning in plant disease detection and classification, suggesting that fine-tuning existing models and exploring new architectures can substantially improve accuracy across various plant species. By drawing on these developments, this research seeks to leverage the strengths of ViTs to tackle the specific challenges of sugarcane leaf disease detection. Despite these advancements, challenges such as dataset diversity, class imbalance, and the variability of real-world agricultural images have not been fully addressed by existing ViT-based approaches in sugarcane disease detection. Therefore, there is a need to investigate the effectiveness of ViTs in sugarcane leaf disease classification and to determine whether they can outperform traditional CNN-based methods.

This study addresses this gap by systematically evaluating the performance of ViTs compared to CNNs in classifying sugarcane leaf diseases, highlighting the advantages and limitations of each approach.This research advances previous work by applying Vision Transformers to sugarcane leaf disease detection and demonstrating their superior performance over traditional CNNs. By fine-tuning a pre-trained ViT model on a comprehensive and augmented dataset of sugarcane leaf images, we aim to overcome the limitations of CNNs in modeling long-range dependencies and handling variations in image scale and orientation. Furthermore, we address issues related to dataset diversity and class imbalance, enhancing the model's generalization and robustness.

The findings of this research contribute to the existing body of knowledge by demonstrating the applicability of ViTs in sugarcane disease detection and addressing challenges related to dataset variability and model generalization. The proposed approach offers a reliable and scalable solution for early disease detection, ultimately supporting better crop management and improving agricultural productivity.

2 Materials and methods

2.1 Dataset

2.1.1 Dataset source and description

The Sugarcane Leaf Disease Dataset utilized in this study was obtained from the Dataset. It contains 19,926 images of sugarcane leaves categorized into six distinct classes, representing various conditions of sugarcane leaves, including healthy and diseased varieties. To enhance model training, the dataset was augmented using rotation, flipping, zooming, resizing, and cropping techniques.

The Six Classes:

1. Bacterial Blight Disease (4,800 images):
   - Sugarcane leaves with water-soaked lesions that become necrotic due to bacterial blight infection.
2. Healthy Leaves (3,132 images):
   - Images of healthy sugarcane leaves with no signs of disease or damage.
3. Mosaic Disease (2,772 images):
   - Leaves exhibiting patchy discoloration forming a mosaic pattern caused by mosaic disease.
4. Red Rot Disease (3,108 images):
   - Sugarcane leaves infected with red rot, a fungal disease characterised by reddening of the stalk and visible leaf symptoms.
5. Rust Disease (3,084 images):
   - Leaves showing brown or orange rust-like spots resulting from fungal infections.
6. Yellow Disease (3,030 images):
   - Leaves displaying yellowing and chlorosis symptoms associated with yellow disease.

2.1.2 Data preprocessing

To ensure compatibility with the ViT model and improve the efficiency of training, the following preprocessing steps were applied:

- Image Resizing: All images were resized to 224 × 224 pixels, matching the input size required by the ViT model.
- Normalization: Pixel values were normalized with a mean of 0.5 and a standard deviation of 0.5 to standardize the dataset.

2.2 Data preprocessing and augmentation

2.2.1 Data preprocessing

To ensure compatibility with the ViT model and improve the efficiency of training, the following preprocessing steps were applied:

Image Resizing: The images were resized to 224 × 224 pixels to conform to the ViT model's input size requirements.

Normalization: To standardize the dataset, pixel values were normalized with a mean of 0.5 and a standard deviation of 0.5.

2.2.2 Data augmentation

Various data augmentation techniques were performed which enhance the model's generalization capabilities and minimize overfitting:

Random Rotation: Images were rotated up to 90 degrees. This simulates different orientations of leaves in real-world scenarios, accounting for varying angles at which leaves might be photographed.

Horizontal and Vertical Flips: Images were flipped horizontally and vertically to mimic the natural variability in leaf orientation, as leaves can be present in various directions in the field.

Zooming: Applied zoom transformations ranging up to 20%. This simulates varying distances between the camera and the leaves, helping the model become invariant to scale changes. Cropping: Random cropping of 80% of the original image size was performed. This approach enables the model to attend to different regions of the leaf and learn to identify disease features, even when only a portion of the leaf is visible. Random Affine Transformations: Adjusted rotation, scaling, and translation for added diversity. This helps the model become robust to geometric transformations that can occur due to camera movement or varying perspectives.

Rationale for choosing specific augmentation techniques:

These specific data augmentation techniques were chosen to simulate real-world variations that occur during image capture in agricultural settings. Factors such as different camera angles, distances, lighting conditions, and partial occlusions are common in field conditions. By integrating these variations into the training data, the model learns to recognize disease patterns in diverse and realistic scenarios, thereby enhancing its ability to generalize to new, unseen data.

2.3 Data splitting

The dataset was divided into the following subsets for model training and evaluation:

- Training Set: Consisting of 70% of the dataset, totalling, 13,948 images.
- Validation Set: Making up 15% of the dataset, which is 2,989 images.
- Test Set: Also comprising 15% of the dataset, amounting to 2,989 images.

Stratified sampling was employed to ensure balanced class distributions across all subsets. A random seed of 42 was used to maintain reproducibility.

2.4 ViT architecture

The ViT model was utilized for the classification of sugarcane leaf diseases. The detailed workflow is as follows:

1. Input Image Preprocessing:
    - Input sugarcane leaf images were resized to 224 × 224 pixels.
    - Each image was divided into 16 × 16 patches, resulting in 196 patches per image.
2. Patch Embedding:
    - Each patch was flattened into a vector and passed through a linear projection layer to create patch embeddings.
3. Positional Encoding:
    - To preserve spatial information, positional encodings were added to the patch embeddings.
4. Transformer Encoder:
    - The Transformer encoder processes the embeddings through:
        - Multi-Head Self-Attention: Models global relationships between patches.
        - Layer Normalization and Feedforward Layers: Enhance feature learning and stabilization.
5. Classification Head:
    - A classification token ([CLS]) was prepended to the patch sequence.
    - To predict one of the six disease classes—Bacterial Blights, Healthy, Mosaic, Red Rot, Rust, and Yellow—the output from the [CLS] token was processed through a Multi-Layer Perceptron (MLP) head.
6. Training Details:
    - The ViT model was trained on a labelled dataset using supervised learning.
    - To enhance generalization, data augmentation techniques—including rotation, flipping, and scaling—were applied. Cross-entropy loss was employed as the objective function.

The architecture diagram (Fig. 1) illustrates the detailed workflow of the ViT model for sugarcane leaf disease prediction.

2.4.1 ViT model

The ViT base model, google/vit-base-patch16-224-in21k, pre-trained on the ImageNet-21 k dataset, was fine-tuned for the task of classifying sugarcane leaf diseases.

Justification for choosing ViT over traditional CNNs:

Although traditional CNNs are widely employed for image classification tasks, they have limitations in capturing long-range dependencies within images due to their localised receptive fields. ViTs utilize a self-attention mechanism, enabling them to grasp global context and relationships across the entire image. This is particularly advantageous for plant disease detection, as the disease manifestations may involve subtle textures and patterns spread over different regions of a leaf. Recent systematic reviews [17] have emphasized the emerging importance of Vision Transformers (ViTs) in plant disease detection tasks. These models have demonstrated superior performance in capturing complex patterns compared to traditional CNNs, especially in datasets where disease symptoms are subtle and spatially distributed [17].

2.4.2 Practical considerations for ViT

Modelling Complex Patterns: ViTs can effectively capture intricate patterns and non-local relationships within images, which is essential for distinguishing between diseases with similar visual symptoms.

Flexibility with Input Data: ViTs require less architectural customization for different input sizes compared to CNNs, providing flexibility when dealing with images of varying resolutions.

Reduced Inductive Bias: ViTs rely less on the inductive biases inherent in CNNs (such as translation invariance and locality), which can be advantageous when such assumptions do not hold in complex datasets like those involving plant diseases with varied manifestations.

Scalability with Transfer Learning: Fine-tuning a ViT pre-trained on a large dataset leverages learned representations, which is beneficial when the available dataset is relatively small, as in specialized agricultural applications.

Given these considerations, the ViT model was chosen over traditional CNNs to potentially achieve better performance and robustness in sugarcane leaf disease classification.

Model specifications:

Input Patch Size: 16 × 16 pixels.

Embedding Dimension: 768.

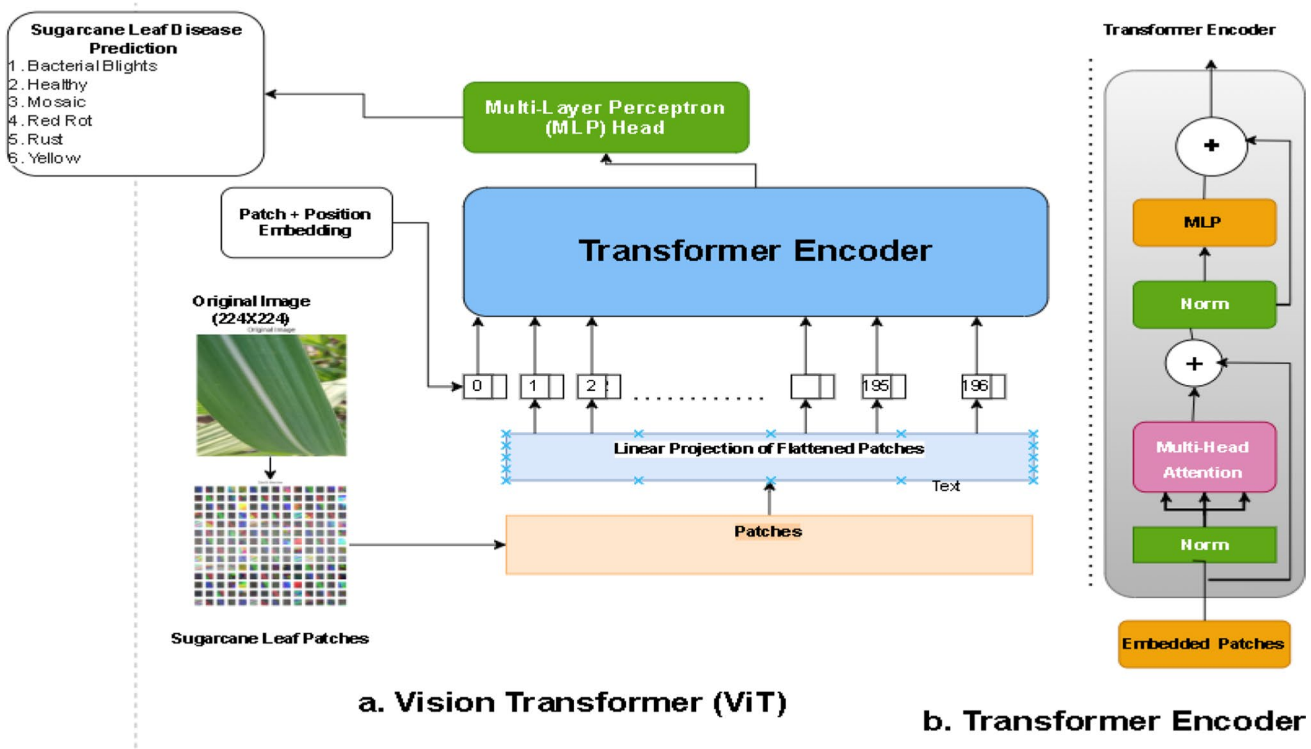


Fig. 1 **a** Architecture diagram of ViT **b** detailed structure of the transformer encoder

Number of Transformer Layers: 12.

Number of Attention Heads: 12.

The final classification head of the model was replaced with a custom layer designed to output probabilities for six classes corresponding to the different sugarcane leaf diseases.

2.4.3 Training parameters

The model training was carried out using the following parameters:

- Optimizer: Adam optimizer with a learning rate of 0.0001.
- Loss Function: Cross-Entropy Loss, suitable for multi-class classification.
- Batch Size: 8, selected to accommodate GPU memory constraints.
- Epochs: 10.
- Learning Rate Scheduler: A Step LR scheduler was employed, reducing the learning rate by a factor of 0.1 every 7 epochs.

2.4.4 Environment setup

Programming Language: Python 3.11.

- Framework: PyTorch 1.9.0
- Libraries:
  - torchvision for data handling and transformations
  - transformers for ViT implementations
  - tqdm for progress visualization during training
- Hardware: NVIDIA GPUs with at least 11 GB VRAM
- Parallelism: Utilized Data Parallelism for efficient computation

2.5 Comparative models selection

To evaluate the performance of the ViT model, two established CNN architectures, ResNet50 and VGG16, were selected as baseline models for comparison.

ResNet50

ResNet50 [18] is a 50-layer deep CNN introduced by He et al. [18]. It incorporates residual learning through skip connections, which mitigates the vanishing gradient problem in deep networks. ResNet50 has demonstrated high performance on various image recognition tasks and serves as a standard benchmark in computer vision research.

VGG16

VGG16 [19], developed by Simonyan and Zisserman [19], is a 16-layer CNN known for its simple and uniform architecture using small 3 × 3 convolutional filters. Despite its simplicity, VGG16 has achieved excellent results on standard image classification datasets and is widely adopted as a baseline model due to its effectiveness in feature extraction."

Rationale for selection

- Benchmarking Performance: Both ResNet50 and VGG16 are renowned for their strong performance in image classification tasks [18, 19]. Including these models provides a solid benchmark to evaluate the effectiveness of the ViT model in comparison to traditional CNN architectures.

- Architectural Diversity: ResNet50 employs deep residual learning, enabling the training of deep networks, while VGG16 utilizes a deep but uniform layer structure. Comparing the ViT model with these architectures allows for assessing how transformer-based models perform relative to CNNs with different design philosophies.
- Relevance to Plant Disease Classification: A recent comprehensive review [17] analyzed 160 studies on deep learning for plant disease detection and highlighted the continued effectiveness of CNNs such as ResNet50 and VGG16 in accurately identifying plant diseases across various datasets. These models have consistently achieved high performance, particularly in classification tasks involving leaf images. The review also notes a growing trend toward using Vision Transformers (ViTs), which offer improved performance in capturing complex, spatially distributed disease patterns. By including both traditional CNNs and transformer-based models in this study, we align with current research directions and benchmark the proposed model against widely accepted architectures in the field [17].

2.6 Evaluation metrics

Several metrics and techniques were employed to assess the model's performance:

- Accuracy: Accuracy is the ratio of correctly predicted instances to the total number of instances, providing an overall measure of model performance.
- Confusion Matrix: Confusion Matrix was.to evaluate the model's performance across individual classes, highlighting any potential biases or misclassifications.
- Visualization:
  - Training and Validation Curves: The training process is monitored to identify any signs of overfitting or underfitting, the training and validation loss and accuracy over the epoch's curves are plotted as shown in the training and validation curves
  - Sample Predictions: Visual comparisons between predicted and true labels to qualitatively evaluate the model's predictions.

3 Results

3.1 Training and validation performance

This section presents the detailed training and validation performance of the ViT model. The results are compared with two traditional CNN architectures: ResNet50 and VGG16 (Table 1).

3.1.1 ViT model results

Training and validation loss and accuracy per epoch

Table 2 shows the training loss, validation loss, training accuracy, and validation accuracy for each epoch during the ViT model training. This detailed breakdown provides insights into the model's learning progression over time and its ability to generalize to unseen data.

Observations from Table:

- Consistent Decrease in Training Loss: The training loss steadily decreases with each epoch, indicating that the model is effectively learning from the training data.
- Improvement in Validation Loss: The validation loss also shows a decreasing trend, suggesting that the model is improving its performance on unseen data and generalizing well.
- High Training Accuracy: The training accuracy increases progressively, reaching 99.80% by the 10 th epoch, which demonstrates the model's strong ability to fit the training data.
- High Validation Accuracy: The validation accuracy mirrors the training accuracy's upward trend, achieving 97.72% by the final epoch, indicating excellent generalization performance.

Table 1 Summary of sugarcane leaf disease dataset

Class	Number of images
Bacterial Blight Disease	4,800
Healthy Leaves	3,132
Mosaic Disease	2,772
Red Rot Disease	3,108
Rust Disease	3,084
Yellow Disease	3,030
Total	19,926

Table 2 Training and validation metrics per epoch for ViT model (batch size 8, learning rate 0.0001, 10 epochs)

Epoch	Training loss	Validation loss	Training accuracy (%)	Validation accuracy (%)
1	0.0482	0.0984	98.29	96.38
2	0.0342	0.0961	98.84	96.55
3	0.0267	0.0939	99.18	96.85
4	0.0248	0.0882	99.23	97.15
5	0.0175	0.0832	99.56	97.39
6	0.0140	0.0812	99.69	97.49
7	0.0145	0.0806	99.59	97.52
8	0.0119	0.0818	99.69	97.32
9	0.0095	0.0802	99.79	97.66
10	0.0088	0.0806	**99.80**	**97.72**

The bolded values indicate the best training and validation accuracy achieved at the final epoch of training, demonstrating peak model performance

- Minimal Overfitting: The small and consistent gap between training and validation accuracies suggests that overfitting is minimal, and the model maintains its performance across both datasets.

Test set performance

After training, the ViT model was evaluated on the test set consisting of 2989 samples:

- Test Loss: 0.1125
- Test Accuracy: 96.53%

These results confirm that the model maintains high accuracy when predicting completely unseen data, further validating its robustness and effectiveness.

3.1.2 ResNet50 and VGG16 model results

Training and validation loss and accuracy per epoch

Table 3 presents the training loss, validation loss, training accuracy, and validation accuracy for each epoch during the ResNet50 model training.

Table 4 presents the training loss, validation loss, training accuracy, and validation accuracy for each epoch during the VGG16 model training (Table 5).

Test set performance

- ResNet50:

  - Test Accuracy: 91.92%.
  - Observations: The model achieved high training accuracy but showed signs of overfitting, as validation accuracy plateaued earlier.

- VGG16:

  - Test Accuracy: 92.30%.
  - Observations: While VGG16 performed well, it converged more slowly and reached lower final accuracy than ViT.

These results indicate that ViT outperforms both ResNet50 and VGG16 in terms of accuracy and generalization ability.

Table 3 Training and validation metrics per epoch for ResNet50 model

Epoch	Training loss	Validation loss	Training accuracy (%)	Validation accuracy (%)
1	0.5137	0.208	67.62	68.43
2	0.3235	0.2853	84.39	84.72
3	0.2732	0.2653	87.05	86.65
4	0.2531	0.2067	88.9	88.9
5	0.2252	0.1871	89.83	89.03
6	0.2075	0.1828	90.73	90.74
7	0.2063	0.2048	91.49	90.9
8	0.1304	0.1456	92.18	91.65
9	0.109	0.1036	93.07	92.01
10	0.1	0.1157	92.16	92.08

Table 4 Training and validation metrics per epoch for VGG16 model

Epoch	Training loss	Validation loss	Training accuracy (%)	Validation accuracy (%)
1	0.5537	0.248	81.62	85.43
2	0.3835	0.2553	88.39	90.72
3	0.2832	0.2053	90.02	89.63
4	0.2631	0.1967	91.07	89.9
5	0.2452	0.1771	91.83	90.03
6	0.2175	0.1728	92.73	90.74
7	0.2078	0.1648	92.49	91.9
8	0.1855	0.1149	93.18	92.65
9	0.169	0.1066	95.07	92.80
10	0.11	0.1057	93.33	92.91

Table 5 Classification report of the sugarcane leaf disease model

Class	Precision	Recall	F1-score	Support
Bacterial blights	1.00	0.99	0.99	720
Healthy	0.95	0.97	0.96	469
Mosaic	1.00	0.93	0.96	415
Red rot	0.98	0.97	0.98	466
Rust	0.97	0.97	0.97	462
Yellow	0.93	1.00	0.96	268
Macro Avg	0.97	0.97	0.97	2800
Weighted Avg	0.97	0.97	0.97	2800

3.1.3 Learning curves

The training and validation loss curves over 10 epochs for the ViT, ResNet50, and VGG16 models are presented in Fig. 2.

Observations:

- ViT Model:
  - Training Loss: Decreased steadily from 0.4503 in epoch 1 to 0.0459 in epoch 10.
  - Validation Loss: Decreased from 0.2489 to 0.0977, plateauing slightly after epoch 7.
- ResNet50 Model:
  - Training Loss: Decreased from a higher initial value, showing a steady decline across epochs.
  - Validation Loss: Decreased but plateaued earlier than the ViT model, indicating potential overfitting.
- VGG16 Model:
  - Training Loss: Showed a gradual decrease, but with a slower convergence rate when compared to ViT.
  - Validation Loss: Plateaued after epoch 6, suggesting limited improvement with additional training.

Analysis:

- The ViT model demonstrates the most significant reduction in both training and validation loss, indicating effective learning and better generalization to the validation data.
- The ResNet50 and VGG16 models exhibit earlier plateauing of validation loss, which may indicate that they reached their optimal performance earlier and gained limited benefits from further epochs.
- The consistent decline in the ViT model's validation loss past epoch 7 suggests that it continues to learn meaningful patterns without overfitting.

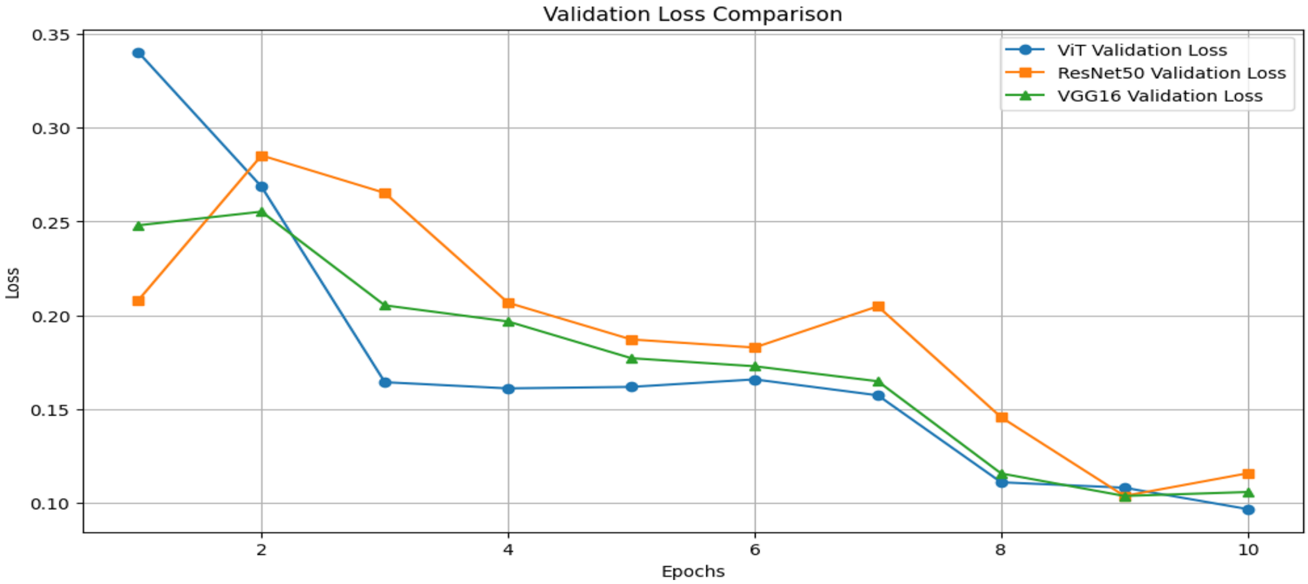


Fig. 2 Training and validation loss curves for ViT, ResNet50, and VGG16

3.1.4 Accuracy curves

Figure 3 illustrates the training and validation accuracy over the epochs for the ViT, ResNet50, and VGG16 models.

Observations:

- ViT Model:
  - Training Accuracy: Increased from 98.29% in epoch 1 to 99.80% in epoch 10.
  - Validation Accuracy: Improved from 96.38% to 97.72% over the same period.
- ResNet50 Model:
  - Training Accuracy: Showed steady improvement but started from a lower initial accuracy compared to ViT.
  - Validation Accuracy: Plateaued around 92% after epoch 7, indicating limited generalization gains with further training.
- VGG16 Model:
  - Training Accuracy: Increased gradually but remained lower than that of the ViT model throughout the epochs.
  - ValidationAccuracy: Plateaued around 93% from epoch 6 onwards.

Analysis:

- The ViT model exhibits the highest training and validation accuracies across all epochs, indicating superior learning capacity and generalization ability.
- The convergence of training and validation accuracies in the ViT model without significant divergence suggests minimal overfitting.

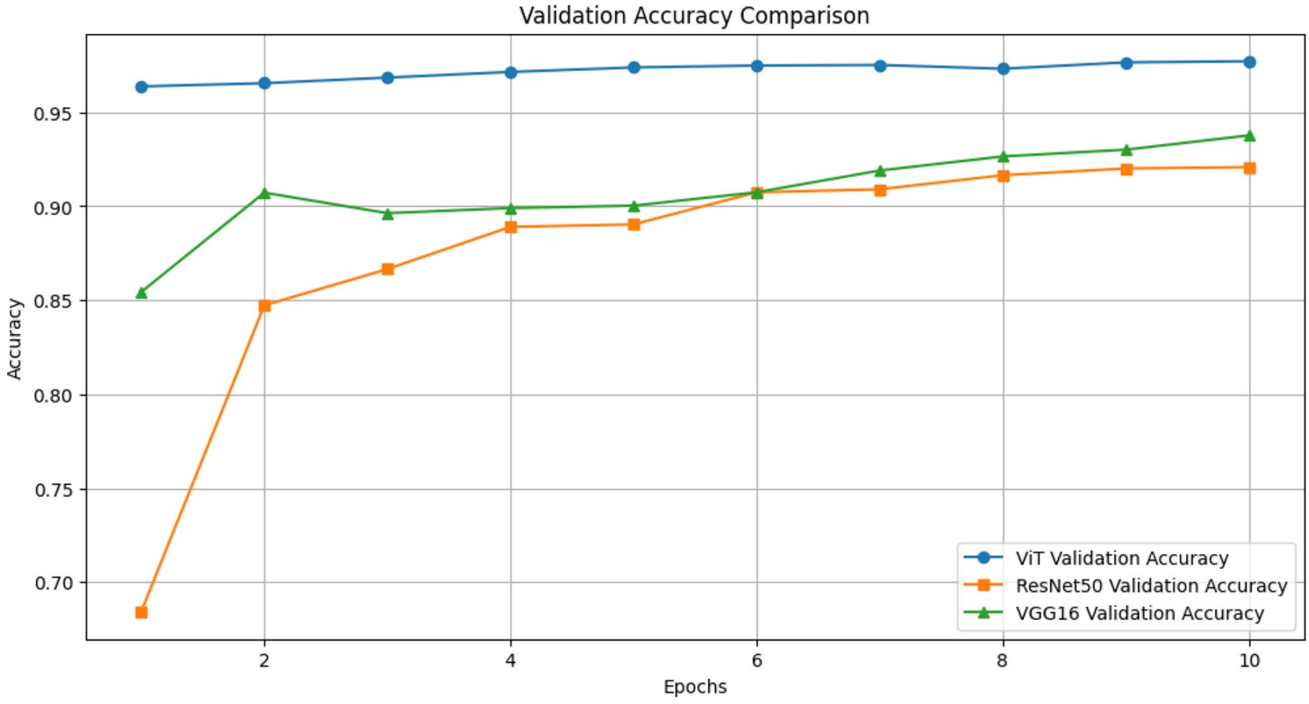


Fig. 3 Training and validation accuracy curves for ViT, ResNet50, and VGG16

- The ResNet50 and VGG16 models show a smaller gap between training and validation accuracies, but their overall accuracies are lower compared to the ViT model.
- The plateauing of validation accuracy in ResNet50 and VGG16 may point to a limitation in their ability to learn additional features beyond a certain point with the given data and training setup.

3.2 Test set performance

3.2.1 Overall accuracy

On the test set comprising 2,994 samples, the ViT model achieved an accuracy of 96.53%, correctly classifying 2,885 samples and misclassifying 104 samples. This high accuracy demonstrates the model's strong performance in classifying sugarcane leaf diseases.

3.2.2 Confusion matrix

The confusion matrix in Fig. 4 illustrates the model's predictions versus true labels. The model performed exceptionally well in classes such as Healthy and Red Rot, with accuracies exceeding 98%. Some confusion occurred between Smut and Rust, likely due to similar visual features.

3.3 Sample predictions and misclassifications

Figure 5 presents sample images from the test set with their predicted and true labels. The model correctly identified various diseases, demonstrating practical applicability.

Observations:

- The model correctly identified various diseases, demonstrating its practical applicability.
- Misclassified samples were analyzed to understand the limitations of the model.

3.3.1 Misclassified samples analysis

The model resulted in a total of 104 misclassified samples during evaluation.

- Misclassifications often occurred between classes with similar visual symptoms.
- Further data augmentation and collection could help the model distinguish these subtle differences better.

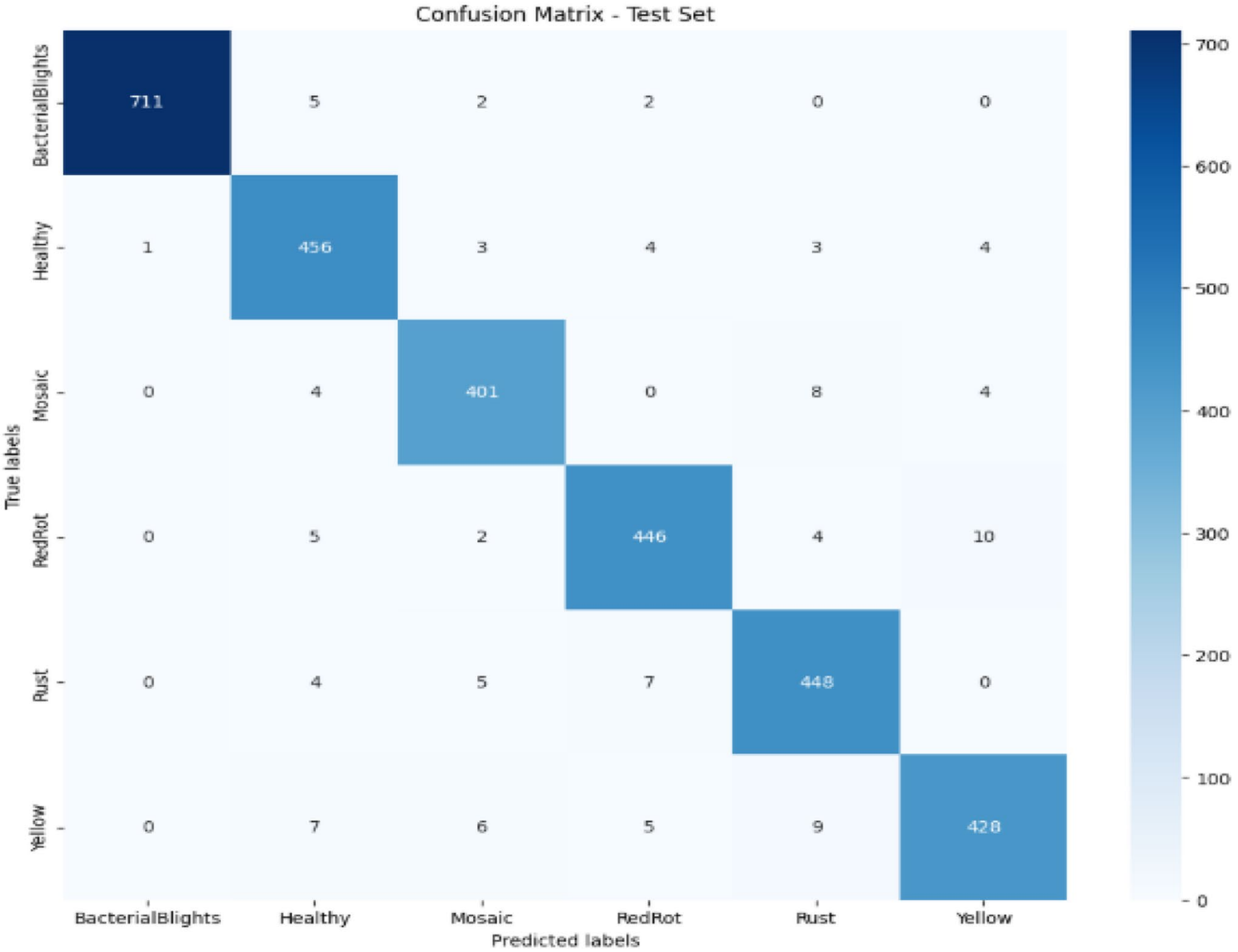
Figure 6 illustrates examples of misclassified sugarcane leaf images, displaying the true disease labels alongside the model's predicted labels.

3.4 Performance evaluation

This section presents a detailed classification report highlighting the performance metrics achieved by the ViT model. The metrics include precision, recall, F1-score, and overall accuracy for each class. These results illustrate the model's exceptional performance and reliability in detecting and classifying diseases effectively (Fig. 7).

3.5 Hyperparameter tuning and performance results

To enhance the performance of the ViT model for sugarcane leaf disease detection, we conducted a series of experiments by adjusting key hyperparameters, including the number of epochs, batch size, learning rate, and optimizer. The results of these experiments are summarized in Table 6.



	BacterialBlights	Healthy	Mosaic	RedRot	Rust	Yellow
BacterialBlights	711	5	2	2	0	0
Healthy	1	456	3	4	3	4
Mosaic	0	4	401	0	8	4
RedRot	0	5	2	446	4	10
Rust	0	4	5	7	448	0
Yellow	0	7	6	5	9	428

Fig. 4 Confusion matrix

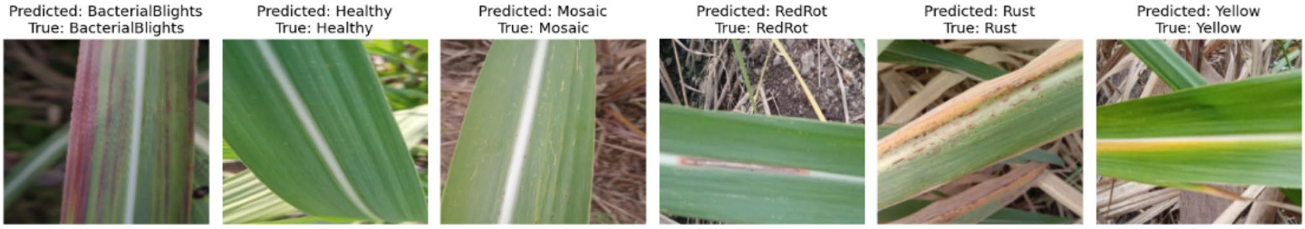


Fig. 5 Sample predictions with true and predicted labels

3.5.1 Analysis of results:

- Experiment No. 7 yielded the best performance, achieving a training accuracy of 99.80% and a test accuracy of 97.72%.
- This configuration used:
  - Number of Epochs: 10
  - Batch Size: 8
  - Learning Rate: 0.0001
  - Optimizer: Adam

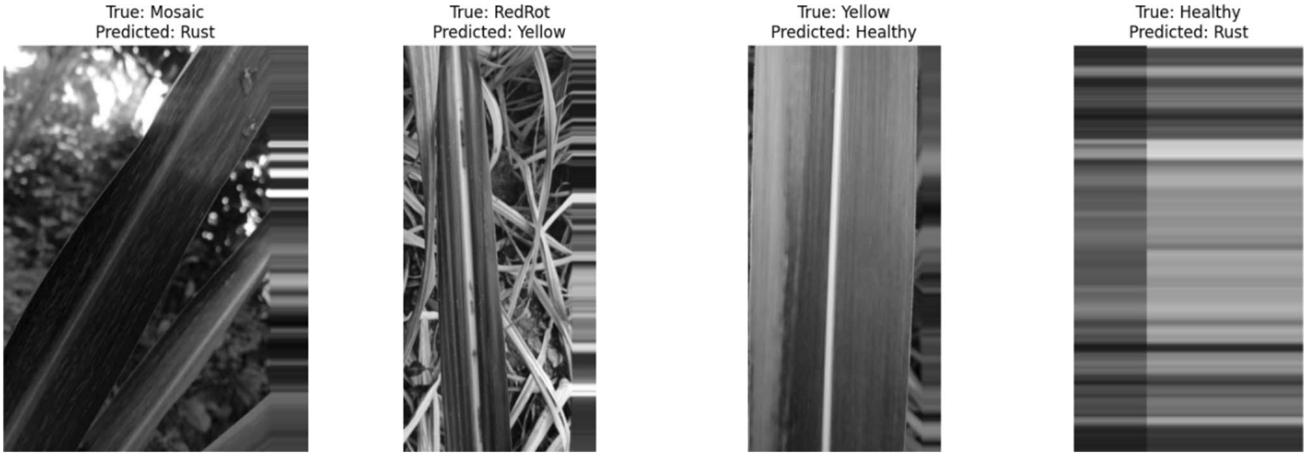


Fig. 6 Misclassified sugarcane leaf images: true vs. predicted labels

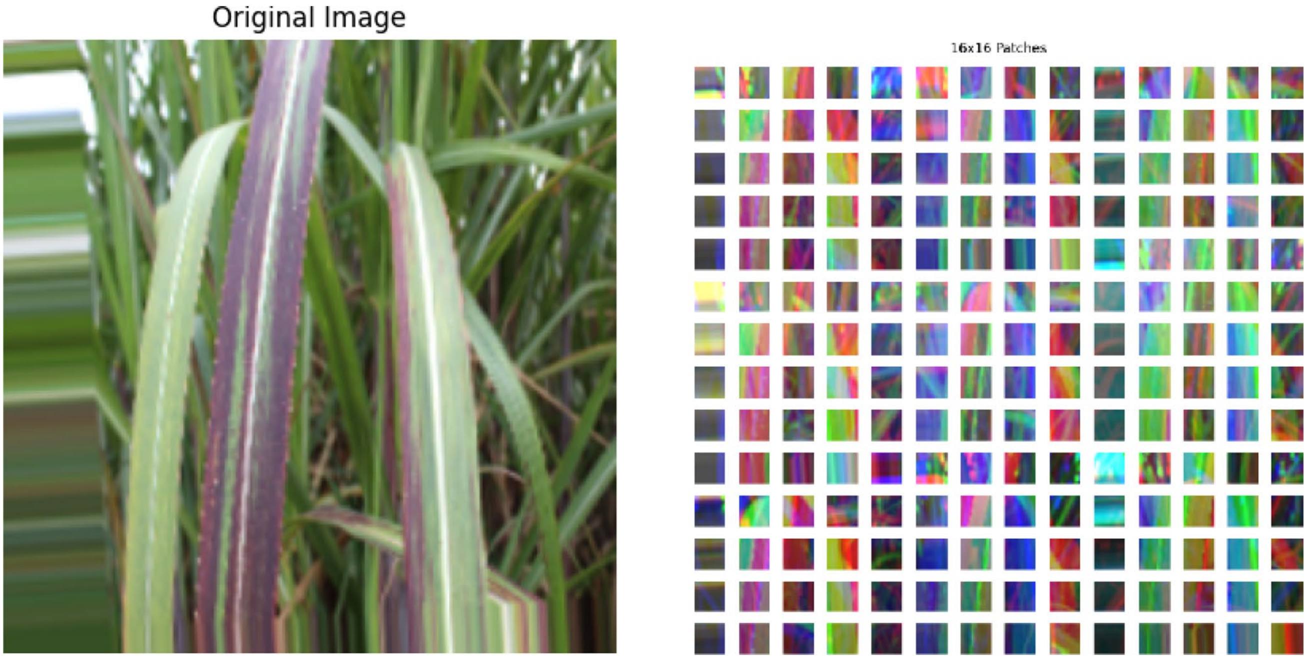


Fig. 7 Original image and its 16 × 16 patches

Impact of batch size:

- Smaller Batch Sizes Improve Accuracy:
  - Reducing the batch size from 32 (Experiment No. 1) to 16 (Experiment No. 2) significantly improved both training and test accuracies.
  - Further reducing the batch size to 8 (Experiment No. 7) led to the highest accuracies observed in our experiments.
- Trade-offs with Training Time and Memory Constraints:
  - Training Time: Smaller batch sizes resulted in longer training times per epoch due to fewer samples being processed simultaneously. For instance, training with a batch size of 8 took approximately twice as long per epoch compared to a batch size of 16.

Table 6 Summary of Hyperparameter Configurations and Model Performance

No	No. of epochs	batch size	Learning rate	Training accuracy (%)	Test accuracy (%)	Optimizer
1	10	32	0.0001	88	86	Adam
2	10	16	0.0001	98.33	96.77	Adam
3	10	16	0.001	58	55	SGD
4	5	32	0.000017	80	78	Adam
5	5	32	0.00017	53	51	Adam
6	10	16	0.0001	94	91	SGD
7	**10**	**8**	**0.0001**	**99.80**	**97.72**	**Adam**
8	10	8	0.001	89.01	88.15	Adam
9	10	8	0.0001	93.74	82.15	SGD
10	10	8	0.001	79.12	75.93	SGD
11	10	64	0.0001	95.54	94.47	Adam

The bolded row represents the optimal hyperparameter configuration (batch size = 8, learning rate = 0.0001, optimizer = Adam) that resulted in the highest training and test accuracy among all experiments

  - Memory Consumption: While smaller batch sizes reduce the memory required per batch, the overall GPU memory usage remains significant due to the model's size. Conversely, larger batch sizes (e.g., batch size 64 in Experiment No. 11) can lead to out-of-memory errors if the GPU does not have sufficient VRAM.

Effect of learning rate and optimizer:

- Learning Rate:
  - A learning rate of 0.0001 proved optimal across different batch sizes when using the Adam optimizer.
  - Increasing the learning rate to 0.001 (Experiments No. 3, 8, and 10) led to reduced accuracies, suggesting that the model failed to converge properly at higher learning rates.
- Optimizer:
  - The Adam optimizer consistently outperformed SGD in our experiments.
  - When using SGD (Experiments No. 3, 6, 9, and 10), the model achieved lower accuracies compared to using Adam with the same hyperparameters.

Impact of epochs:

- Increasing the number of epochs from 5 (Experiments No. 4 and 5) to 10 improved model performance, indicating that more training iterations allowed the model to learn better representations.

3.5.2 Time and memory constraints

Training iime:

- Batch size impact:
  - Batch Size 8**:** Training took significantly longer per epoch due to processing fewer samples at a time. For example, with a batch size of 8, each epoch took approximately 17 min.
  - Batch Size 16: Training time per epoch was around 8–9 min.

- Batch Size 32 and 64: Training times per epoch were 5 min and 4.27 min however, larger batch sizes did not always lead to better performance.

3.5.3 Conclusion from hyperparameter tuning

The hyperparameter tuning experiments underscore the importance of selecting appropriate values for batch size, learning rate, optimizer, and number of epochs. Configuration 2 emerged as the optimal setting for this study, balancing the various factors to achieve the highest accuracy.

Based on these findings, the ViT model underwent final training, emphasizing that meticulous hyperparameter tuning is vital for enhancing the performance of deep learning models in agricultural disease detection tasks.

3.6 Visualization of image patches

To understand how the ViT model processes images, we visualized the transformation of an image into 16 × 16 patches.

4 Discussion

4.1 Interpretation of results

The ViT model demonstrated high accuracy in classifying sugarcane leaf diseases, indicating its capability to learn complex patterns from images. ViTs, utilize self-attention mechanisms to capture global contextual information, which may contribute to their enhanced performance over CNNs in this context.

4.2 Comparison with previous studies

Previous studies utilizing CNNs for sugarcane leaf disease classification reported accuracies ranging between 90 and 95% [6]. In comparison, the ViT model proposed in this study demonstrated superior performance, achieving a test accuracy of 96.77% These results demonstrate the benefits of transformer-based architectures. Their ability to capture global contextual information through self-attention mechanisms makes them a promising alternative to traditional CNNs for agricultural image classification tasks.

4.3 Limitations

The encouraging results of this study suggest several directions for future research to enhance the application of ViTs in agricultural disease detection.

4.4 Future work

4.4.1 Integration with aerial imagery for large-scale monitoring

Future research could explore integrating the ViT model with aerial imagery from drones or satellites to enable large-scale monitoring of sugarcane crops. High-resolution aerial images can facilitate the detection of diseases across extensive agricultural areas, providing timely information for disease management. Adapting the model to process aerial imagery will involve addressing challenges like varying resolutions and perspectives.

4.4.2 Dataset expansion and diversity

Expanding the dataset to include images from different regions, environmental conditions, and growth stages will improve the model's robustness. Incorporating real-world field images with variations in lighting and backgrounds will enhance the model's ability to generalize to practical applications.

4.4.3 Model optimization for edge devices

Optimizing the ViT model for deployment on mobile devices or edge computing platforms can provide farmers with accessible, real-time diagnostic tools. Techniques such as model pruning and quantization can reduce computational requirements, making the model suitable for resource-constrained environments.

4.4.4 Multi-label disease classification

Developing the model to handle cases where leaves are affected by multiple diseases simultaneously will increase its applicability. Implementing multi-label classification algorithms will enable the detection of co-occurring diseases.

4.4.5 Collaboration with agricultural experts

Engaging with plant pathologists can enhance data annotation quality and ensure the model focuses on agriculturally significant features. Expert insights will aid in refining the model and improving its relevance to farming practices.

5 Conclusion

This study demonstrates the effectiveness of ViTs in classifying sugarcane leaf diseases, achieving superior performance compared to traditional CNN-based approaches. By leveraging the self-attention mechanisms of ViTs, the model effectively captures intricate patterns in sugarcane leaf images, resulting in high training and validation accuracy. The model's strong performance on test data, evidenced by an accuracy of 96.53%, highlights its promise as an effective tool for detecting and managing diseases in agriculture. The proposed methodology, which includes robust data preprocessing, augmentation techniques, and fine-tuning of a pre-trained ViT model, addresses many challenges associated with traditional disease detection methods, such as inefficiency, subjectivity, and scalability issues. However, limitations such as dataset size and potential class imbalances highlight the need for further research to improve model robustness and applicability to real-world scenarios.

Future directions include expanding the dataset to incorporate diverse environmental conditions, exploring hybrid architectures, and developing practical tools for field deployment. By integrating state-of-the-art machine learning techniques with agricultural practices, this study contributes to sustainable farming solutions, enabling timely interventions and improved crop productivity.

Author contributions S.M. (Saritha Miryala) conceived and designed the study, collected and prepared the dataset, performed data preprocessing and data augmentation, implemented and fine-tuned the Vision Transformer (ViT) model, conducted the experiments, and analyzed the results. S.M. also wrote the initial draft of the manuscript and prepared all the figures and tables. K.R. (Dr. Krupa Rasane) provided guidance and expert advice on the research methodology and deep learning techniques, contributed to the interpretation of the findings, and critically reviewed and revised the manuscript for important intellectual content. Both authors contributed to the manuscript revision, read, and approved the final version for publication.

Funding This research did not receive a specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability The datasets generated during and/or analysed during the current study are available in the Kaggle repository, accessible at [https://www.kaggle.com/datasets/akilesh253/sugarcane-plant-diseases-dataset].

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Singh A, Yadav V. A review of image processing techniques for plant disease detection. Agric Inform J. 2022;12(3):45–52.
2. Kaur A, Gupta S. Automating potato leaf disease detection with lightning-fast CNNs: Precision using PyTorch Lightning. Proceedings of the 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), IEEE. 2024.
3. Ghosh R, Roy S. Applications of deep learning in agriculture: A case study on sugarcane leaf disease classification. Proceedings of the International Conference on Smart Agriculture Technologies, 101–110. Springer. 2023.
4. Shruthi U, Nagaveni V, Raghavendra BK. A review on machine learning classification techniques for plant disease detection. Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 281–284. IEEE. 2019.
5. Saleem MH, Potgieter J, Arif KM. Plant disease detection and classification by deep learning. Plants. 2019;8(11):468. https://doi.org/10.3390/plants8110468.
6. Demilie WB. Plant disease detection and classification techniques: A comparative study of performances. J Big Data. 2023;10: Article 73. https://doi.org/10.1186/s40537-023-00773-5.
7. Islam MR, Rajon MSI, Hasan HMM, Siddik MAB. Smart web application for plant leaf disease detection using CNN and pre-trained models. Comput Electron Agric. 2023;210:107936. https://doi.org/10.1016/j.compag.2023.107936.
8. Shoaib M, Shah B, El-Sappagh S, Ali A, Ullah A, Alenezi F, Gechev T, Hussain T, Ali F. An advanced deep learning models-based plant disease detection: a review of recent research. Front Plant Sci. 2022;13: Article 902112. https://doi.org/10.3389/fpls.2022.902112.
9. Dosovitskiy A, Beyer L, Kolesnikov A et al. An image is worth 16×16 words: Transformers for image recognition at scale. Proceedings of the International Conference on Learning Representations (ICLR 2021). 2021.
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017). 2017; 5998–6008. Curran Associates Inc.
11. Borhani Y, Khoramdel J, Najafi E. A deep learning-based approach for automated plant disease classification using Vision Transformer. Sci Rep. 2022;12: Article 19151. https://doi.org/10.1038/s41598-022-23072-7.
12. Boukabouya RA, Berrimi M, Moussaoui A. Vision Transformer-based models for plant disease detection and diagnosis. Proceedings of the 2022 5th International Symposium on Informatics and its Applications (ISIA), IEEE. 2022.
13. Barman U, Sarma P. ViT-SmartAgri: vision transformer and smartphone-based plant disease detection for smart agriculture. Agronomy. 2024;14(2):327. https://doi.org/10.3390/agronomy14020327.
14. Kunduracıoğlu I, Paçal I. Data-efficient vision transformer models for robust classification of sugarcane leaf diseases. J Soft Comput Decis Anal. 2024;2(1):258–71. https://doi.org/10.31181/jscda21202446.
15. Kunduracioglu I. Utilizing ResNet architectures for identification of tomato diseases. J Inf Digit Manag. 2024;1(1):1–10. https://doi.org/10.59543/jidmis.v1i.11949.
16. Kunduracioğlu I, Paçal I. Advancements in deep learning for accurate classification of grape leaves and diagnosis of grape diseases. J Plant Dis Prot New Ser. 2024. https://doi.org/10.1007/s41348-024-00896-z.
17. Pacal I, Kunduracioglu I, Alma MH, Deveci M, Kadry S, Nedoma J, Slany V, Martinek R. A systematic review of deep learning techniques for plant diseases. Artif Intell Rev. 2024;57: Article 304. https://doi.org/10.1007/s10462-024-10944-7.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). 2016; 770–778. IEEE. https://doi.org/10.1109/CVPR.2016.90
19. Food and Agriculture Organization of the United Nations (FAO), International Fund for Agricultural Development (IFAD), United Nations Children's Fund (UNICEF), World Food Programme (WFP), & World Health Organization (WHO). The state of food security and nutrition in the world 2023. Geneva: FAO; 2023.