Hybrid CNN - ViT Framework for Guava Leaves and Fruits Disease Detection in Precision Agriculture

Nebras Jalel Ibrahim[1], Farah Hatem Khorsheed[1], Zainab Hassan Mohammed[1], Alaulddin Mueen Latfa[2], Walaa Badr Khudhair[1] , Zahraa Fadhil Hassan[3] and Hussein Alkattan[4]
[1]*Computer Center, University of Diyala, 32001 Diyala, Iraq*
[2]*Baghdad Agriculture Directorate, 10001 Baghdad, Iraq*
[3]*Department of Quality Assurance and University Performance, University of Diyala, 32001 Diyala, Iraq*
[4]*Department of System Programming, South Ural State University, 454080 Chelyabinsk, Russia*
{nebras.jale l, farah_hatam}@ uodiyala.edu.iq, {eng.zaynab.hassan.2025, alaamueen, walaabadr2023, zahraa1998fadhil, alkattan.hussein92}@gmail.com

Keywords: Guava Disease Detection, Convolutional Neural Networks (CNN), Vision Transformers (Vit), Hybrid Deep Learning, Image Classification, Precision Agriculture, Plant Disease Detection.

Abstract: The detection of guava plant diseases is essential for ensuring fruit quality and reducing agricultural losses. In this study, we developed a hybrid deep learning model that integrates Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) to classify guava leaf diseases from image data. The model was trained and validated on an augmented dataset comprising five categories: Disease Free, Phytophthora, Red Rust, Scab, and Styler and Root. Experimental results demonstrate that the proposed hybrid model achieved an overall accuracy of 96%, with a macro-averaged F1-score of 0.97. Class-level performance showed near-perfect detection for Disease Free and Red Rust, while slightly lower but still robust results were obtained for Styler and Root (F1-score = 0.93). The confusion matrix indicates strong generalization across all classes with minimal misclassification. Despite the strong classification metrics, highlighting the need for further optimization in probability calibration. Overall, the results confirm that combining CNN feature extraction with ViT's sequence modeling provides an effective strategy for guava disease recognition, offering a reliable decision-support tool for precision agriculture.

1 INTRODUCTION

Plant diseases have a major effect on agricultural productivity, resulting in decreased yields, huge financial losses, and deterioration of crop quality. Guavas (Psidium guajava) are unique among tropical fruits because they are high in dietary fiber and vitamin C, which improve nutrition and food security in many regions of the world. Guava production is susceptible to a variety of leaf and fruit diseases, such as Phytophthora, Scab, Red Rust, Styler, and Root, which can significantly lower the fruit's market value and shelf life after harvest [1]. Because of these difficulties, early and precise disease detection is a crucial component of sustainable horticulture. Disease diagnosis Guava Orchards has traditionally depended on visual observation by farmers or agricultural experts. This method is not only subjective but time-consuming and highly dependent on the level of expertise of the individual carrying out the inspection. Manual methods can lead to a misidentification problem and delays in taking timely measures that further exacerbate the infection intervention. With large-scale guava orchards rapidly increasing, it has become increasingly impractical to rely on human inspections thereby inspiring artificial intelligence (AI) and computer vision techniques to be integrated for automating the detection process [2], [3]. Automated methods eliminate dependency on manual observation ensuring consistency, scalability, and fast response which is much needed in a precision agriculture system. Deep Learning strategies have emerged as very powerful techniques toward the identification of diseases in plants over recent years [4]. Convolutional Neural Networks (CNNs) exhibit an exemplary performance in feature extraction through spatial and textual features of images relating to leaves and fruits for a robust classification mechanism against diseases of various crops [5]. However, the principal architecture of

CNNs basically emphasizes local pattern features (analogous to spots, edges, or discolorations) and hence fail to clearly articulate long-range dependency as well as a holistic image context [6]. The weakness becomes more glaring in the recognition of guava disease since its symptoms manifest themselves in irregular or diffused patterns over the surface area of a leaf.

A Vision Transformer (ViTs) has recently been proposed as an alternative paradigm in the field of computer vision. Whereas CNNs learn relationships among patches, the self-attention mechanism of ViTs relates features at arbitrary positions over the whole image and thus integrates global and local pattern information on disease manifestation conditions [7], [8]. So far, ViTs proved useful for several domains concerning image classification and object detection tasks in agricultural imagery since they allowed capturing efficiently the patterns of disease development, interference from the background, and structural variability of plant leaves.

Noting the filling strengths in between CNNs and ViTs, there is currently a trend in the research whereby hybrid architectures involve an amalgamation of both local feature extractors and global context modelers into one unified framework [9]. Such integration will ensure a balanced approach: fine-grained lesion characteristics can be picked up by CNNs, and long-range spatial dependencies within the whole leaf can be acquired by ViTs. This synergy promises boosted classification accuracy, robustness, and generalization in guava disease detection toward a strong move for pointing the future direction toward powerfulness in next-generation precision agriculture systems [10].

This paper introduces a CNN-ViT hybrid model for classifying guava diseases. Data employed covers both original and augmented images of guava leaves representing five different categories of diseases besides the healthy class. The major contributions that have resulted from this study are:

- Development of a hybrid architecture by merging local spatial learning, as provided by CNN, with global sequence modeling through ViT.
- Test the model on enhanced data of Guava leaves and Fruits Disease, and present classification results that are at par with the best in the industry.
- Give an elaborate performance analysis using the parameters of accuracy, precision, recall, F1-score, confusion matrix.

This establishes hybrid deep learning as a very appropriate methodology toward increasing the accuracies of recognizing diseases in Guava plants for timely recognition and precision agriculture practices.

2 RELATED WORK

Several recent research have looked into guava disease detection using machine learning and deep learning approaches, indicating an increasing interest in automating plant pathology.

Rashid et al. [11] proposed a hybrid deep learning model of a MobileNetV2-U-Net segmentation block and a YOLOv5 detector, which produced a result of 92.41% accuracy of segmentation and a mean average precision of 71% for several diseases of guava leaves, such as anthracnose and wilt. Doutoum et al. (2023) [12] also compared a few pre-trained CNNs and it resulted that EfficientNet-B3 performed best on guava leaf datasets for 94.93% accuracy supporting the benefits of transfer learning for tiny samples. Jain et al. [13] proposed a unified CNN model for multi-class guava fruit disease detection, achieving 95.90% accuracy, which outperformed several single-disease detection methods., Tewari et al. [14] dataset and reported ~95% accuracy, but generalization across circumstances was limited. Hashan et al. [15] focused practical application, using CNNs into horticulture imaging systems for guava fruit recognition. Mumtaz et al. [16] used a hybrid deep learning framework to detect and analyze leaf blight in the guava plants. Their model combines ResNet-50 for deep feature extraction with a traditional machine learning classifier to enhance accuracy and interpretability. Using a dataset of custom-captured guava leaves, the system obtained a classification accuracy of 99.3%, showing the effectiveness of combining convolutional features with a lightweight classifier in the recognition of guava diseases. The study emphasizes practical deployment by balancing performance with computational efficiency. Yashu et al. [17] blended CNNs and Random Forests to show how hybrid techniques might increase interpretability in guava disease diagnosis. In another practical area, Nandi et al. [18] developed a lightweight CNN optimized for mobile deployment, allowing for resource-efficient disease detection on both leaves and fruits.

More recent work has focused on detection and management. Shetty et al. [19] evaluated YOLOv5 and Faster R-CNN for disease detection in multiple crops, including guava and mango, and found that

YOLO excelled in localization tasks in real-time conditions. Chouhan et al. [20] used a CNN-Random Forest hybrid to diagnose and manage guava illness, demonstrating the benefits of merging deep learning and ensemble approaches. Paramesha et al. [21] conducted a comparison of machine learning and deep learning strategies for guava disease classification and found that DL approaches consistently outperformed ML in terms of accuracy and scalability. Finally, Nasra et al. [22] developed a deep CNN-based pipeline for multi-class guava disease detection, with reporting accuracies more than 95%, demonstrating the maturity of CNN techniques for automated detection in guava agriculture. Table 1 shows the literature on the Guava Leaves and Fruits Disease detection.

Overall, these papers show constant development in guava illness detection, transitioning from single-disease CNN models to hybrid deep learning pipelines, lightweight architectures, and object detection frameworks. While most achieve excellent performance (83-96%), there are still issues with multi-disease identification, real-time deployment, and strong generalization in field situations.

3 METHODOLOGIES

The proposed methodology consists of five stages: dataset preparation, preprocessing, hybrid CNN-ViT architecture, training, and evaluation metrics.

3.1 Dataset Description

The Guava Leaves and Fruits Dataset, which is publicly available at [23], was used for the experimental evaluation in this study. In order to guarantee precise labeling and a range of disease representation, this dataset was gathered under the expert supervision of the Bangladesh Agricultural University in Mymensingh, Bangladesh. Through the use of deep learning and computer vision techniques, the dataset seeks to support research on automated guava disease detection and recognition.

Table 1: Summary of literature on the guava leaves and fruits disease detection.

Year	Authors	Method / Model	Dataset	Crop Focus	Key Results
2023	Rashidet al.[11]	Hybrid Deep Learning: GIP-MU-Net (MobileNetV2 encoder + U-Net decoder) for lesion segmentation, and YOLOv5-based model for multi-disease detection on leaves	Guava fruit images	Guava	92.41% segmentation accuracy on average for infected patches;
2023	Doutoumet al. [12]	Transfer Learning with CNNs: compared four pre-trained CNN models; EfficientNet-B3 performed best	1,834 guava leaf images in 5 classes (canker, dot, mummification, rust, healthy)	Guava (leaf).	94.93% accuracy
2023	Jain et al. [13]	an integrated CNN-based approach for multi-class guava disease classification (combining features for a single model)	Guava fruit disease image dataset (multiple disease types; exact size not stated)	Guava (fruit) diseases	95.90% accuracy
2023	Tewari et al. [14]	CNN variants	Custom dataset	Guava	~95% accuracy
2023	Hashan et al. [15]	CNN	Horticulture dataset	Guava	Practical deployment
2023	Mumtaz et al. [16]	Hybrid deep learning: ResNet-50 for feature extraction + ML classifier	Custom guava leaf dataset (captured via imaging)	Guava (leaf blight)	using a hybrid approach combining ResNet and ML models
2023	Yashu et al. [17]	CNN + Random Forest	Field dataset	Guava	Enhanced interpretability
2023	Nandi et al. [18]	Lightweight CNN	Fruit + leaf dataset	Guava	Mobile-friendly deployment
2024	Shetty et al. [19]	YOLOv5, Faster R-CNN	Multi-crop dataset	Guava & Mango	Strong localization
2024	Chouhan et al. [20]	CNN + Random Forest	Experimental dataset	Guava	Diagnosis & management
2025	Paramesha et al. [21]	ML + DL hybrid	Comparative analysis	Guava	DL outperformed ML
2025	Nasra et al. [22]	Deep CNN	Multi-class dataset	Guava	Automated detection 95%

To ensure the robustness and generalization of the trained models, the dataset includes high-resolution RGB images (512 × 512 pixels) of guava leaves and fruits in a variety of lighting, orientation, and background conditions. There are two subsets offered: the augmented dataset, which has about 4,899 images produced by transformations like rotation, scaling, translation, and flipping, and the original dataset, which has about 527 images. By preventing overfitting, this augmentation enhances model performance.

There are five different classes to which each image in the dataset belongs:

- Disease-Free. Guava fruits and leaves are healthy and show no outward signs of illness.
- Phytophthora. Phytophthora parasitica-caused brown necrotic lesions are seen in infected samples.
- Red rust is distinguished by the presence of orange or reddish pustules on the leaf surface.
- Scab. Displaying tiny, dark, round lesions that could combine to form asymmetrical scabby patches.
- Styler and Root Rot. Usually caused by a fungal infection, this condition manifests as decay at the fruit's styler end or root area.

Seventy percent of the dataset was used for training, fifteen percent for validation, and fifteen percent for testing. To prevent bias during model learning, each subset keeps the distribution of classes balanced. To improve the model's capacity for generalization, all images were resized to 224 × 224 pixels, normalized, and then exposed to additional real-time augmentation (rotation, brightness adjustment, and horizontal flipping) before training.

Both convolutional and transformer-based hybrid models, like the CNN-ViT architecture suggested in this study, can be evaluated using this dataset since it offers a trustworthy standard for evaluating vision-based deep learning architectures.

Figure 1 shows different types of diseased and healthy Guava Leaves and Fruits from this dataset. This preprocessing approach enabled efficient training while preserving key visual disease features [24].

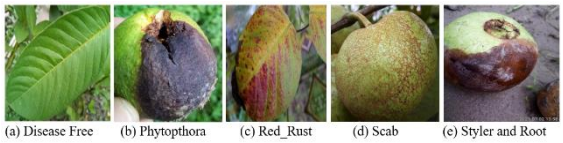

Figure 1: Different samples of the Guava Leaves and Fruits Dataset.

3.2 Preprocessing

To ensure consistency, minimize computational complexity, and improve model generalization, multiple preprocessing techniques were implemented to the Guava Leaves and Fruits Dataset prior to training the hybrid CNN-ViT model:

3.2.1 Image Resizing

Images were scaled to 64 × 64 pixels with RGB color channels, resulting in a uniform input size for both CNN and ViT branches.

3.2.2 Normalization

To stabilize training and accelerate convergence, all pixel values in the guava leaf and fruit images were rescaled to the range [0, 1]. This was carried out using the following formula:

$$x' = \frac{x}{255}. \quad (1)$$

Where:

- x is the original pixel intensity in the range [0, 255];
- $x' \in [0, 1]$ is the normalized pixel value.

3.2.3 Data Augmentation

To alleviate class imbalance and strengthen the model, standard data augmentation approaches were used on both leaf and fruit images. Augmentation operations included:

- Random rotation of up to ±20° to imitate diverse leaf and fruit orientations.
- Flipping horizontally and vertically to match natural leaf placement.
- Random cropping to imitate partial visibility of disease symptoms.
- Brightness correction for field-collected photos.

3.2.4 Label Encoding

The dataset's categorical labels (Phytophthora, Red Rust, Scab, Styler & Root, and Healthy) were encoded as one-hot vectors to ensure compatibility with the model's softmax output layer.

3.2.5 Dataset Splitting

To guarantee class balance across all subsets, the dataset was divided into training (70%), validation

(15%), and testing (15%) subsets using a stratified technique.

All preprocessing procedures were used to guarantee consistency and enhance generalization before to feeding the dataset into the hybrid CNN-ViT model.

4 HYBRID CNN-VIT MODEL

By using the best parts of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs), the proposed model can reliably classify diseases in guava leaves and Fruits. ViTs are great at modeling long-range dependencies and global contextual information using self-attention mechanisms. CNNs, on the other hand, are great at capturing local spatial details like leaf textures, edges, and color variations. By using both local and global feature representations and combining these two designs, the hybrid model makes illness detection more accurate and general. Figure 2 shows an overview of the suggested Hybrid CNN-ViT Model framework.

These modifications increased intra-class variability, reduced the likelihood of overfitting, and enabled the model to learn more generic illness traits.

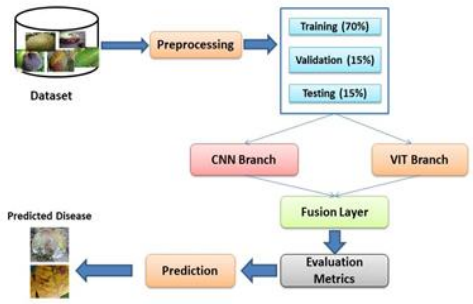

Figure 2: Workflow of the proposed Hybrid CNN-ViT model framework for guava leaves and fruits disease detection.

4.1 CNN Branch

The proposed hybrid model's CNN branch is made to effectively extract local spatial information from photos of guava leaves and Fruits, which are essential for precise disease classification. To gradually lower spatial dimensions while maintaining prominent features, it is composed of three convolutional layers with 32, 64, and 128 filters, each succeeded by a max-pooling layer. To reduce the number of parameters and prevent overfitting, a global average pooling layer is used to create a fixed-size feature vector. After that, a dense layer consisting of 256 units and a dropout rate of 0.3 is applied to the retrieved features, producing reliable and broadly applicable depictions of leaf and Fruit properties. This CNN-based feature extractor serves as the basis for integration with the Vision Transformer in the hybrid CNN-ViT model by efficiently capturing hierarchical, disease-specific patterns including textures, edges, and lesions. The definition of each convolution operation is:

$$f_{i,j}^{(k)} = \sigma(\sum_{m=1}^{M}\sum_{n=1}^{N} w_{m,n}^{(k)}.\, x_{i+m,j+n} + b^{(k)}), \quad (2)$$

where:

- $f_{i,j}^{(k)}$= feature map at location (i,j) for filter k.
- $w_{m,n}^{(k)}$= kernel weights.
- xi+m,j+n = input pixel value.
- b(k) = bias term.
- σ (.)= ReLU activation.

Pooling was applied using:

$$p_{i,j} = \max_{(m,n) \in R} f_i + m, j + n, \quad (3)$$

4.2 Vision Transformer (ViT) Branch

In this proposed hybrid CNN-ViT model, the Vision Transformer (ViT) branch assists the CNN branch in local feature extraction by capturing global contextual linkages across the guava leaf and Fruit pictures. In order to improve classification accuracy for intricate disease patterns, the ViT branch uses self-attention processes to represent long-range dependencies and spatial interactions between various leaf and Fruit regions. Self-attention is used by the ViT branch to capture global contextual information.

4.2.1 Patch Embedding

Input image $X \in R^{H \times W \times C}$ is split into N patches of size $P \times P$:

$$X_p \in R^{N \times (P^2 \cdot C)}, \quad (4)$$

- H: Input image height (pixels). In this study, H=64. W: Input image width (pixels). In this study.
- W=64. C: Number of color channels (depth). For RGB images, C=3.
- P: Patch size (pixels). P=16 in this implementation.
- N: Number of patches. Calculated as $N = (H/P) * (W/P)$. For 64 * 64 images and 16 * 16 patches, N=16.

4.2.2 Linear Projection and Position Encoding:

$$Z_0 = [x_{cls}; X_p]W_E + E_{pos}. \quad (5)$$

Where:

- W_E is the projection matrix;
- x_{cls} is the class token;
- E_{pos} is the positional embedding.

4.2.3 Multi-Head Self Attention (MHSA):

The MHSA mechanism transforms the input sequence of tokens $Z \epsilon R^{(N+1)*D}$ *into queries (Q), keys (K), and values (V):*

$$Attention(Q,K,V) = softmax\left(dk\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (6)$$

with

$$Q = XW_Q,\ K = XW_K,\ V = XW_V, \quad (7)$$

where:

- W_Q, W_K, W_V: Learned weight matrices used to project the input tokens into the query, key, and value spaces, respectively.
- d_K: The dimension of the keys (and queries) for a single attention head. This is calculated as d_K=*D*/heads.
- heads: The number of parallel attention mechanisms (heads), set to heads=4.
- D: Projection dimension (Hidden size). The dimensionality of the token embedding, set to D=64.

4.2.4 Feed-Forward Network (FFN):

$$Z' = LayerNorm(X + MHSA(X)), \quad (8)$$

$$Z = LayerNorm\big(Z' + FFN(Z')\big). \quad (9)$$

4.3 Fusion and Classification

The fusion mechanism combines the local, low-level features F_{CNN} extracted by the CNN branch (a 256-dimensional vector) with the global, contextual features F_{ViT} extracted by the ViT branch (a 256-dimensional vector from the class token output), allowing the model to use both global contextual information and local spatial details for precise guava leaf and Fruits disease classification [25], [26].

The hybrid CNN-ViT model achieves high classification accuracy and robust F1-scores across all guava disease categories because to the Fusion Layer, which makes sure the model captures both fine-grained features and holistic leaf and Fruits structure. Features from ViT and CNN are combined:

The fusion operator used in this work is vector concatenation (denoted by the symbol $\|$), which simply merges the two feature vectors into a single, comprehensive feature vector *F*:

$$F = [F_{CNN} \parallel F_{ViT}], \quad (10)$$

this concatenated feature vector is then passed to a dense layer for final classification. The final prediction $\acute{Y}i$ for class *i* is obtained using the Softmax activation function:

$$\acute{Y}i = \frac{e^{zi}}{\sum_{j=1}^{K} e^{zi}}, \quad (11)$$

where Zi is the input score (logit) for class *i*, K=5 is the number of classes.

4.4 Training

The suggested hybrid model's training configuration and hyperparameter values are compiled in Table 2. These setups maintained balanced learning across all classes, avoided overfitting, and guaranteed steady convergence.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} y_{i,} \operatorname{clog} y\hat{}_{i,c}\,. \quad (12)$$

Table 2: Training configuration and hyperparameters of the hybrid CNN-ViT model.

Parameter	Description
Device	HP 250 G8 Notebook PC, Intel® Core™ i5-1135G7 @ 2.40GHz, 8 GB RAM, Intel® Iris® Xe Graphics
Frameworks	TensorFlow 2.14.0, Keras 2.14.0, Python 3.10
Epochs	100
Batch Size	32
Optimizer	Adam
Initial Learning Rate	0.001
Learning Rate Schedule	Step decay (factor = 0.1 on validation plateau)
Early Stopping	Enabled (patience = 10 epochs)
Weight Decay (L2)	1e-5
Class Weighting	Applied to balance underrepresented classes

4.5 Evaluation Metrics

Several evaluation metrics were used to assess the performance of the proposed hybrid CNN-ViT model for guava leaf disease classification. The evaluation included accuracy, precision, recall, and F1-score in order to provide a comprehensive assessment of the

classification performance across all disease categories. Accuracy was used to measure the overall classification performance of the model, while precision and recall were employed to evaluate the reliability of positive predictions and the capability of the model to correctly identify diseased samples. In addition, the F1-score was used to provide a balanced evaluation between precision and recall, particularly in the presence of class imbalance.

5 RESULTS

The effectiveness of the suggested hybrid architecture was evaluated by comparing three models: the hybrid CNN-ViT model, a Vision Transformer (ViT), and a pure CNN. To ensure fairness in comparison, all models were trained using the same preprocessing pipeline and data splits. Confusion matrix analysis, F1-score, recall, classification accuracy, and precision were the main evaluation criteria.

5.1 Training and Validation Performance

The model showed stable convergence, as shown by the training and validation curves (see Fig. 3). Good generalization without noticeable overfitting was confirmed by the validation accuracy, which reached about 96%, and the validation loss, which slowly reduced and stabilized after multiple epochs.

The CNN model's training and validation accuracy and loss curves are shown in Figure 4. The CNN model's training and validation curves demonstrate a consistent increase in accuracy and a decrease in loss over time. With a classification accuracy of roughly 86%, the model demonstrated strong learning and good generalization, with only slight overfitting.

The Figure 5 demonstrates that the ViT model reached approximately 88% accuracy with little overfitting and good generalization performance, exhibiting a steady increase in accuracy and a consistent decline in loss over epochs.

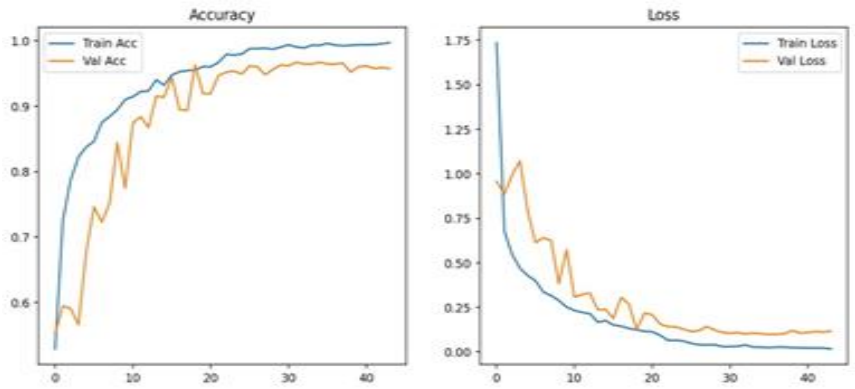

Figure 3: Training and testing performance: Accuracy and loss over epochs for hybrid CNN-ViT model's.

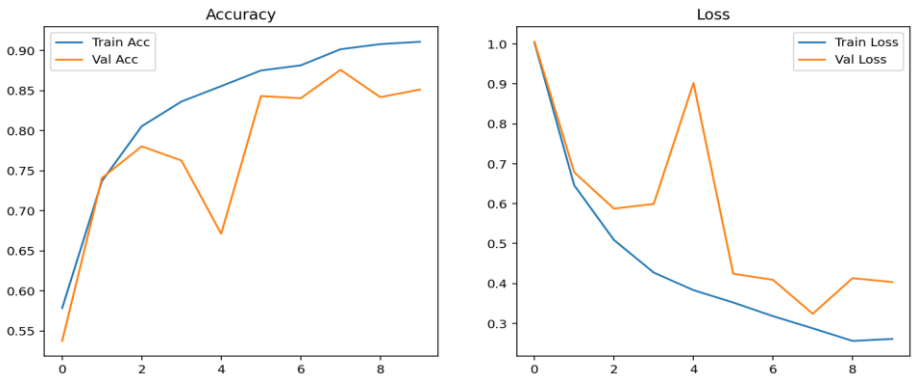

Figure 4: Training and testing performance: Accuracy and loss over epochs for CNN model's.

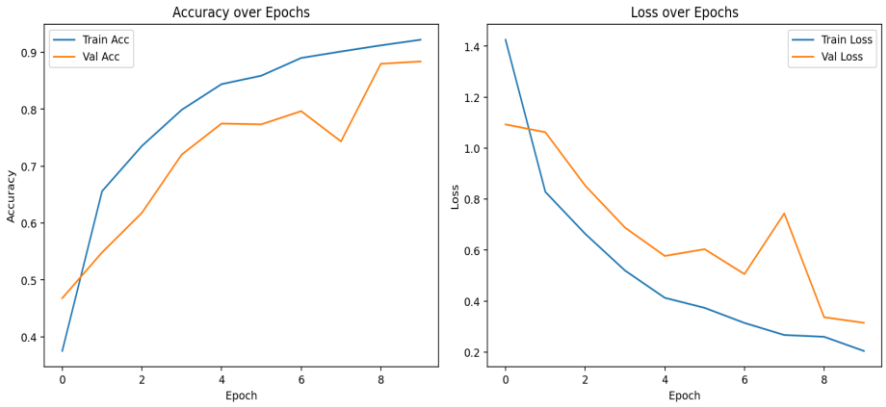

Figure 5: Training and testing performance: Accuracy and loss over epochs for ViT model's.

5.2 Classification Report

Tables 3, 4, and 5 show the F1-score, precision, and recall for each class. For the classification of multi-class plant diseases, three models were compared: There are three models: the CNN-ViT hybrid model, the CNN pure model, and the Vision Transformer (ViT) model. The overall accuracy of the CNN, ViT, and CNN-ViT hybrid models was approximately 86%, 88%, and 96%, respectively.

The CNN-ViT hybrid performed better than the other models in every class, but it was especially good at identifying Scab and Styler & Root. These findings show that the CNN-ViT hybrid offers balanced and extremely accurate classification, which qualifies it for use in actual agricultural disease monitoring systems.

Table 3: Classification performance per class for hybrid CNN-ViT models.

Class	Precision	Recall	F1-score
Disease Free	0.98	0.97	0.98
Phytophthora	0.95	0.96	0.96
Red Rust	0.99	0.98	0.99
Scab	0.96	0.95	0.96
Styler & Root	0.92	0.93	0.93

Table 4: Classification performance per class for CNN model's.

Class	Precision	Recall	F1-Score
Disease Free	0.84	0.95	0.90
Phytophthora	0.92	0.87	0.89
Red Rust	0.98	0.94	0.96
Scab	0.71	0.97	0.82
Styler and Root	0.92	0.62	0.74

Table 5: Classification performance per class for ViT model's.

Class	Precision	Recall	F1-Score
Disease Free	0.97	0.99	0.98
Phytophthora	0.84	0.76	0.80
Red Rust	0.97	0.91	0.94
Scab	0.84	0.91	0.87
Styler and Root	0.80	0.84	0.82

5.3 Confusion Matrix

This study uses confusion matrix analysis to assess the performance of three deep learning models for multi-class plant disease classification: CNN-ViT (Model 1) in Figure 6, CNN (Model 2) as shown in Figure 7, and ViT (Model 3) as shown in Figure 8. Five different classes - Disease Free, Phytophthora, Red Rust, Scab, and Styler & Root - were used to test the models. Styler & Root and Scab were the main areas of misclassifications in CNN's (Model 2) mediocre performance.

ViT (Model 3) had some minor misclassifications but improved detection for multiple classes. The CNN-ViT hybrid model (Model 1), which demonstrated superior robustness and potential for practical deployment in automated plant disease monitoring systems, achieved the highest accuracy (96%) and consistently strong performance across all classes, as summarized in Table 6.

Table 6: Accuracy comparison of deep learning models for plant disease classification.

Model	Accuracy
CNN	0.86
ViT	0.88
Hybrid CNN-ViT	0.96

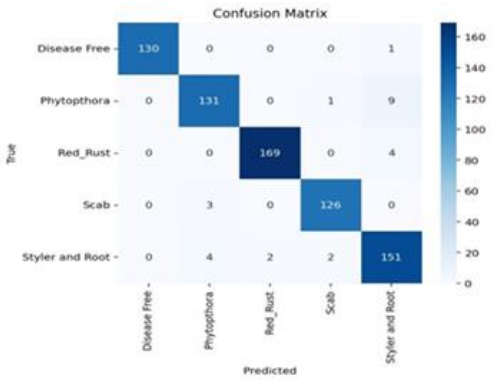

Figure 6: Confusion Matrix of CNN-ViT model.

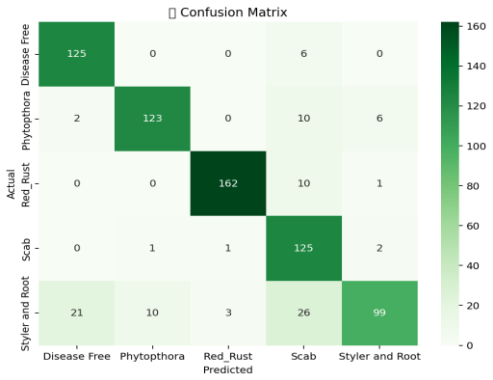

Figure 7: Confusion Matrix of CNN model

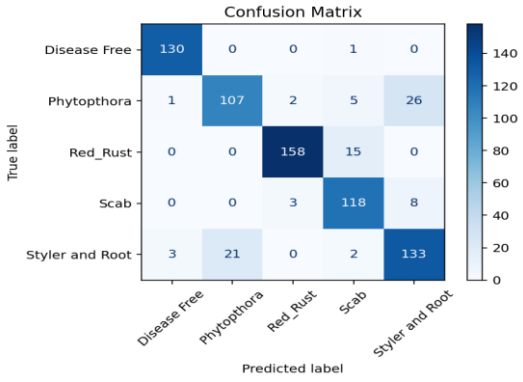

Figure 8: Confusion Matrix of ViT model's

5 CONCLUSIONS

This study introduced a hybrid deep learning framework that combines Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for the classification of diseases affecting guava leaves and fruits. The model showed great performance on the Guava Leaves and Fruits Dataset, with 96% accuracy and a weighted F1-score. It did better than traditional CNN- or SVM-based methods.

The CNN branch did a good job of capturing local textural features like rust spots and scab lesions, while the ViT branch modeled global contextual dependencies, which made it easier to tell visually similar classes apart. The classification report and confusion matrix showed that the proposed method worked well and gave reliable results for most disease categories. However, it is still hard to tell the difference between closely related classes like Scab and Styler and Root.

In general, the hybrid CNN-ViT method has shown to be a useful tool for automatically identifying guava diseases. It can help farmers make better decisions about precision agriculture and may even help them lose less crops in real-world farming situations.

6 FUTURE WORK

Despite the promising results, several directions remain for future improvement. First, future models should support multi-label classification, since guava leaves may contain multiple diseases simultaneously. Second, expanding the dataset with more field-acquired images under varying environmental conditions would improve model generalization and robustness. Third, lightweight architectures such as MobileViT or EfficientNet-Lite should be explored to enable deployment on mobile and edge devices in smart farming environments. In addition, integrating the proposed framework with IoT-based monitoring systems and blockchain technologies may support real-time and traceable agricultural applications. Finally, incorporating explainable AI techniques, such as Grad-CAM and attention visualization, could improve model interpretability and increase user trust in automated disease diagnosis systems.

REFERENCES

[1] A. S. M. F. Al Haque, R. Hafiz, M. A. Hakim, and G. M. R. Islam, “A computer vision system for guava disease detection and recommend curative solution using deep learning approach,” in Proc. Int. Conf. on Computer and Information Technology (ICCIT), 2019. doi: 10.1109/ICCIT48885.2019.9038598.

[2] A. Mehmood, M. Ahmad, and Q. M. Ilyas, “On precision agriculture: Enhanced automated fruit disease identification and classification using a new ensemble classification method,” Agriculture, vol. 13, no. 2, p. 500, 2023. doi: 10.3390/agriculture13020500.

[3] G. Shrestha, M. Das, and N. Dey, "Plant disease detection using CNN," in Proc. 2020 IEEE Applied Signal Processing Conference (ASPCON), Kolkata, India, 2020, pp. 280–284. doi: 10.1109/ASPCON49795.2020.9276755.

[4] J. A. Brenes, M. Eger, and G. Marín-Raventós, "Early detection of diseases in precision agriculture processes supported by technology," in Sustainable Intelligent Systems, Singapore: Springer, 2021, pp. 11–33. doi: 10.1007/978-981-33-6866-6_2.

[5] W. Ding, C. Yu, Y. Liu, and J. Li, "Next generation of computer vision for plant disease monitoring in precision agriculture: A contemporary survey, taxonomy, experiments, and future direction," Information Sciences, vol. 665, p. 120338, 2024. doi: 10.1016/j.ins.2024.120338.

[6] F. H. Khorsheed, R. Hazim, S. A. Hassan, and Q. Saihood, "Hybrid CNN-XGB framework for enhancing human activity recognition," Fusion: Practice & Applications, vol. 15, no. 2, 2024.

[7] P. S. Thakur, A. Kumar, A. Mehta, A. Saxena, and A. Agarwal, "Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT," arXiv preprint arXiv:2207.07919, 2022.

[8] A. Almadhor, A. Alturki, N. Rauf, M. R. H. Siddiqi, A. Alghamdi, M. I. Khan, and F. A. Khan, "AI-driven framework for recognition of guava plant diseases through machine learning from DSLR camera sensor based high-resolution imagery," Sensors, vol. 21, no. 11, p. 3830, 2021. doi: 10.3390/s21113830.

[9] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Computers and Electronics in Agriculture, vol. 145, pp. 311–318, 2018. doi: 10.1016/j.compag.2018.01.009.

[10] F. Alqahtani et al., "A hybrid deep learning model for rainfall in the wetlands of southern Iraq," Modeling Earth Systems and Environment, vol. 9, no. 4, pp. 4295–4312, 2023.

[11] J. Rashid, I. Khan, G. Ali, S. U. Rehman, F. Alturise, and T. Alkhalifah, "Real-time multiple guava leaf disease detection from a single leaf using hybrid deep learning technique," Computers, Materials & Continua, vol. 74, no. 1, pp. 1235–1257, 2023. doi: 10.32604/cmc.2023.032005.

[12] A. S. Doutoum, R. Eryiğit, and B. Tuğrul, "Classification of guava leaf disease using deep learning," WSEAS Transactions on Information Science and Applications, vol. 20, pp. 356–363, 2023. doi: 10.37394/23209.2023.20.38.

[13] R. Jain, P. Singla, Niharika, R. Sharma, V. Kukreja, and R. Singh, "Detection of guava fruit disease through a unified deep learning approach for multi-classification," in Proc. IEEE Int. Conf. Contemporary Computing and Communications (InC4), Bangalore, India, 2023, pp. 305–309. doi: 10.1109/InC457730.2023.10262886.

[14] V. Tewari, N. A. Azeem, and S. Sharma, "Automatic guava disease detection using different deep learning approaches," Multimedia Tools and Applications, vol. 82, 2023. doi: 10.1007/s11042-023-15909-6.

[15] A. M. Hashan et al., "Smart horticulture based on image processing: Guava fruit disease identification," in Proc. IEEE SCOReD, 2023. doi: 10.1109/SCOReD60679.2023.10563573.

[16] Mumtaz, M. I. Malik, M. U. Ghaffar, A. Munir, S. A. Khan, and A. Arif, "A hybrid framework for detection and analysis of leaf blight using guava leaves imaging," Agriculture, vol. 13, no. 3, p. 667, 2023. doi: 10.3390/agriculture13030667.

[17] V. Kukreja, K. Madan, Yashu, A. Singh, and D. Kumar, "Precision agriculture: Guava disease diagnosis via CNN and Random Forest," in Proc. 2023 3rd Int. Conf. on Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 2023, pp. 1–6. doi: 10.1109/SMARTGENCON60755.2023.10442410.

[18] R. N. Nandi, A. H. Palash, N. Siddique, and M. G. Zilani, "Device-friendly guava fruit and leaf disease detection using deep learning," in Proc. Int. Conf. on Machine Intelligence and Emerging Technologies, Springer LNCS, 2023. doi: 10.1007/978-3-031-34619-4_5.

[19] K. U. Shetty et al., "Plant disease detection for guava and mango using YOLO and Faster R-CNN," in Proc. IEEE IATMSI, 2024. doi: 10.1109/IATMSI60426.2024.10503209.

[20] D. Chouhan, M. Kumari, C. Kumar, and V. Kukreja, "From detection to action: Managing guava diseases using CNN and Random Forest models," in Proc. IEEE AUTOCOM, 2024. doi: 10.1109/AUTOCOM60220.2024.10486171.

[21] K. Paramesha, S. Jalapur, S. Hanok, K. Puttegowda, G. Manjunatha, and B. Kumara, "Machine learning and deep learning approaches for guava disease detection," SN Computer Science, vol. 6, no. 1, 2025. doi: 10.1007/s42979-025-03886-6.

[22] O. Güler, T. Etem, and M. Teke, "Hybrid augmentation for multi-channel deep learning in guava leaf disease detection," Ain Shams Engineering Journal, vol. 16, 2025. doi: 10.1016/j.asej.2023.103716.

[23] Mendeley Data, "Guava disease dataset," [Online]. Available: https://data.mendeley.com/datasets/x84p2g3k6z/1.

[24] A. J. Yousif and M. H. Al-Jammas, "Real-time Arabic video captioning using CNN and transformer networks based on parallel implementation," Diyala Journal of Engineering Sciences, pp. 84–93, 2024.

[25] H. Alkattan et al., "The prediction of students' academic performances with a classification model built using data mining techniques," in IET Conf. Proc. CP824, vol. 2022, no. 26, pp. 353–356, 2022.

[26] N. Ibrahim, A. Abbas, and F. Khorsheed, "A systematic review for misuses attack detection based on data mining in NFV," Sakarya University Journal of Computer and Information Sciences, vol. 6, no. 3, pp. 239–252, 2023.