Enhancing Guava Fruit Disease Detection and Localization through a Hybrid Vision Transformer and Convolutional Neural Network Architecture

MSc Research Project
MSc in Data Analytics

Gokul Lala
Student ID: x22222227

School of Computing
National College of Ireland

Supervisor: Abid Yakoob

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Gokul Lala		
Student ID:	x22222227		
Programme:	MSc in Data Analytics	**Year:**	:2023-2024
Module:	MSc Research Project		
Supervisor:	Abid Yakoob		
Submission Due Date:	12/8/2024		
Project Title:	**Enhancing Guava Fruit Disease Detection and Localization through a Hybrid Vision Transformer and Convolutional Neural Network Architecture**		
Word Count:	5991	**Page Count**	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Gokul Lala

Date: 16/09/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Guava Fruit Disease Detection and Localization through a Hybrid Vision Transformer and Convolutional Neural Network Architecture

Gokul Lala
x22222227

Abstract

This research proposes a novel hybrid architecture that combines vision transformers and convolutional neural networks to treat the most difficult problem in guava fruit disease detection and localization. At the bottom of this new approach lies the central idea of exploiting complementary strengths from both architectures to help in efficient agriculture-based disease detection.

In this research, a proposed ViT-CNN-based hybrid model is developed and implemented using PyTorch; the implementation is done over 6,549 guava fruit images across the nine classes of diseases. This is a global feature acquired pre-trained ViT that integrates custom stages of CNN layers for local feature refinement and a classification head.

Results indicate that the model comes out very strong in running up to 95.63% validation accuracy. Some of the notable results include very high per-class F1-scores ranging from 0.88 to 1.00, with good handling of class imbalance and very fast convergence during training. This hybrid approach thus overcame some of the limitations of standalone ViTs in capturing fine-grained features relevant to disease identification.

This work thus offers a very powerful tool for early and accurate guava fruit disease detection, generally contributing much to related areas of computer vision in agriculture. From the experiments, it seems that the model is performing well across most classes of diseases and so may be promising for crop management and yield improvement in guava cultivation. Some future works could include multimodal integration, temporal fitting of the model on disease progression, and adaption to different cultivation conditions.

1 Introduction

Guava is a tropical fruit of immense economic and nutritional value brought down by various disease challenges that hit hard on its quality and yield. Hence, timely and proper detection of these diseases in the fields is important for effective management and sustained productivity. Deep learning approaches, especially Convolutional Neural Networks, had recently shown great promises in plant disease detection. However, recent emergence of vision transformers has opened new possibilities in image analysis tasks such as disease detection.

Although ViTs are quite good at learning global dependencies and long-range interactions, they would often miss the understanding of some specific local spatial features or hierarchies that might have importance in this task. In a complementary manner, CNNs capture these local features effectively but potentially lose the context of broader information. To this effect, this paper proposes the fusion of both models' capabilities into one architecture with the view of increasing accuracy and robustness in guava fruit disease detection and localization.

The primary research question this study addresses is**: How can a hybrid architecture integrating Vision Transformers (ViT) with Convolutional Neural Networks (CNN) improve the accuracy and F1 score of guava fruit disease detection and localization on a comprehensive dataset containing various disease types and severities under diverse imaging conditions?**

To answer this question, we set the following objectives:

1. Designing and implementing an advanced, hybrid ViT-CNN architecture for the classification of guava fruit disease.
2. Compare the performance of the hybrid model with that of only ViT and CNN.
3. Analyze the model's effectiveness across different disease types and severities.
4. Check the model's robustness under different imaging conditions.

The success of these objectives will be measured through:

- Comparative analysis of accuracy and F1 score between the hybrid model and stand-alone models.
- Evaluation of the model's performance on a diverse dataset of guava fruit images.

Methodologically, it will make use of a comprehensive dataset of images of guava fruit showing different pathologies at different levels of severity, acquired under different imaging conditions. In the present study, the hybrid architecture will be implemented in PyTorch using pre-trained ViT models and custom-made layers of the CNN. In this context, we will make use of such advanced training techniques as mixed precision training and learning rate scheduling to ensure that the model will perform at its best. Model performance will be assessed by standard metrics: accuracy and the F1 score, while confusion matrices are used as visual analysis tools.

It serves both academic and agricultural interests. Firstly, it is an academic investigation into how much the Cisco transformer-based and convolutional architectures can provide in terms of synergy in carrying out computer vision tasks. On the agricultural side, this helps in looking out for more accurate and robust disease detection tools eventually, crop management practices will have an improved turnaround in yields.

The remainder of the report is organized as follows: Section 2 discusses related work in hybrid deep learning architectures and plant disease detection. The research methodology is expounded in Section 3, after which Section 4 elaborates on the design specifications of our proposed model. Section 5 describes the details of implementation. In Section 6, a full evaluation is made on our model with experimental results and discussions. Finally, Section 7 concludes this report appropriately with directions for future work.

2 Related Work

2.1 Evolution of Deep Learning in Plant Disease Detection

In the last years, deep learning methods related to plant disease detection have shown outstanding achievement, especially concerning guava plants. The pioneer of this was Howlader et al. (2019) ,proposed a deep CNN method for the identification of diseases in guava leaves. Their model, trained on a custom dataset named BU_Guava_Leaf, was able to classify major diseases in the guava leaf with several diseases, such as Algal Leaf spot, Whitefly, and Rust, with an accuracy of 98.74%. This is an important step forward from the traditional manual inspection method and has showcased the potential for automated AI-driven disease detection in agriculture.

Thangaraj et al. (2023) built on this very foundation with an exhaustive and comparative study of various deep learning models employed for disease detection in guava leaves. Their work assessed state-of-the-art architectures such as DenseNet121, DenseNet169, InceptionV3, and Xception. The research used a Kaggle dataset that included images of guava leaf Canker, Dot diseases, Mummification, and Rust, with the rest consisting of healthy samples. Their results indicated that among these models, DenseNet169 performed the best, at 96.12% accuracy. In general, the main contribution of the study was to focus on highlighting the power of transfer learning and pre-trained models in restricted scenarios concerning dataset size a scenario typical of agricultural applications.

2.2 Emergence and Applications of Vision Transformers in Computer Vision

The idea of Vision Transformers (ViTs) put forward by Dosovitskiy et al. in (2020) largely revised the approaches to computer vision. As Berroukham et al.,(2023) nicely suggested, it is the special ability of ViTs to catch global dependencies and model long-range interactions through self-attention mechanisms, which is a weakness in the functioning of CNNs, considering only local features. The authors also pointed to the flexibility of the ViTs for images of various sizes and, together with strong transfer learning, this means possible improved generalization across various tasks. Pan (2022) further pursued the applications of ViTs in image classification tasks, specifically in penguin classification. In this paper, he compares the efficiency of ViT against recurrent neural networks in various classification tasks, hence proving their versatility. The result showed that general ViT can potentially apply to most computer vision tasks, including highly complex agricultural image analysis.

2.3 Vision Transformers in Plant Disease Detection

The application of ViTs in plant disease detection has shown promising results and pushed boundaries as far as accuracy and efficiency are concerned. Maurya et al. (2023) proposed an automated system for the early detection of plant diseases by employing a fine-tuned ViT architecture. The result, therefore, turned out to be very high, reaching 98.22% on the PlantVillage Dataset, thus proving that such intricate features of leaves could be learned by ViTs in order to identify the disease. The study focused on the need for early disease detection to prevent widespread crop damage, especially in populations that depend on agriculture.

Further efforts on ViTs applied to plant leaf disease classification by Rethik and Singh, (2023), tested three different ViT models and obtained test accuracies of 85.87%, 89.16%, and 94.16%. Notice that this study promoted another aspect of the fact: the ability of ViTs to identify specific places on the leaf, which is very helpful in guiding treatment applications. Spatial localization potential is characteristic of only the ViTs as compared to traditional CNN-based approaches.

2.4 Challenges and Limitations of Vision Transformers

Despite their potential, several concerns for practical applications are raised by ViTs primarily because of resource-constrained settings that are frequently encountered in agricultural scenarios. The application of ViT models has various reported limitations, such as high computational costs and large amounts of training data, according to Mia et al. (2023) and Han et al. (2023). These cited limitations are quite daunting for deploying ViTs on edge devices widely used for Agri-field-based applications.

This is in a way showing that Han et al. further emphasized the need for improved transformer architectures, which are designed in a more orientationally friendly way toward computer vision tasks and more lightweight models that can be used with devices possessing less powerful resources. In other words, addressing such challenges is the key if ViTs are to be adopted widely in practical applications including, most importantly, plant disease detection systems.

2.5 Hybrid Approaches: Combining ViTs and CNNs

Owing to a few limitations of ViTs, and as a quest continues to leverage their advantages, researchers recently investigated hybrid approaches in which a combination of ViTs with CNNs was explored. Lin et al. (2024)proposed a cross-architecture knowledge distillation recipe with teacher collaboration for knowledge transfer from large ViTs to compact CNNs. Their approach aims at bridging the architectural gaps between ViT and CNN by feature reaggregation and logit correction distillation; this has hit top performance on CIFAR-100.

This hybrid approach is in line with the idea of integrating ViTs and CNNs into an architecture that ensures further efficiency and improved performance in most tasks related to computer vision, such as plant disease detection. The power of this kind of hybrid model lies in the fact that, on one hand, it can effectively capture global contextual information via the ViTs and fine-grained local features via the CNNs, which are relevant and very important in the identification of diseases accurately, together with their localization.

2.6 Advanced Techniques in Guava Disease Detection

Recent research has employed state-of-the-art deep learning techniques to push the limits of guava disease detection. Mostafa et al. (2022) proposed a deep CNN with overly sophisticated data augmentation techniques for the comparative identification of diseases in guava plants. Their method has exploited color histogram equalization and unsharp masking to increase the quality and quantity of training data. Five neural network structures were drawn in their study: AlexNet, SqueezeNet, GoogLeNet, ResNet-50, and ResNet-101. Their findings demonstrated an accuracy of 97.74% in classification when ResNet-101 was used.

This work is particularly noteworthy for its broad view on data preprocessing and augmentation. The authors applied nine different angle rotations with a view to bringing diversity into the training dataset to underpin the role of data augmentation in improving model robustness and generalization capability. Lastly, this study is very current, with locally collected data on diseases in guavas from Pakistan, hence bridging this gap by providing regionally specific datasets for agricultural AI applications.

2.7 Future Directions and Research Opportunities

The literature review reveals several promising research directions for improving ViTs and hybrid models in plant disease detection:

1. Efficient ViT Architectures: For computationally constrained agricultural settings, developing lightweight ViT models is the key factor in their adoption. This may be done either by designing lightweight attention mechanisms or by compressing techniques of the models.
2. Multi-task Learning: In another perspective, this research is intended to explore the possibility of using VITs to multitask, like detecting diseases, estimating their severity, and predicting yield, to develop more comprehensive systems for monitoring the health of plants.
3. Data Efficiency:Research in techniques that make ViTs require less data, for example, by using self-supervised or few-shot learning, could increase their applicability in situations where little labeled training data is a common AI-for-agriculture challenge..
4. Domain-specific Adaptations: Models could be attuned to particular agricultural applications, having uniquely designed ViT architectures for special challenges such as lighting and occlusions occurring within a field setting..
5. Interpretability and Explainability:Probably, the development of methods for visualizing and explaining the decision-making process would increase trust in, and hence the adoption of, ViTs and hybrid models among agricultural practitioners..

2.8 Research Gap and Proposed Contribution

While there has been considerable progress applying deep learning techniques to guava disease detection, the full potential of the ViTs is yet to be explored. To this regard, the proposed research is going to fill up that gap through developing a hybrid ViT-CNN architecture for the detection and location of diseases in guava fruits..

This would give the global feature extraction capabilities of ViTs and the strengths in local feature extraction of CNNs, thereby potentially overcoming the limitations of each architecture when used alone. Further work on addressing questions of computational efficiency and more generally in leveraging strengths from both architectures could add significant value to this study in plant disease detection and related applications in the field of computer vision in agriculture.

Such success will make available a more accurate and robust plant disease detection system to farmers, thereby improving the crop management practices of agricultural farmers and reducing losses. Moreover, experiences learned from this hybrid approach will aid the development of similar architectures in further complex vision tasks beyond plant disease detection, thus helping in enriching the broad field of agricultural technology and sustainable farming practices.

3 Research Methodology

In this research, the authors have proposed a hybrid of Vision Transformer and Convolutional Neural Network architecture for the detection and localization of guava fruit diseases. This methodology was designed in such a way to utilize the strengths of both ViTs and CNNs and solve the limitations previously identified in the literature.

3.1 Data Collection and Preparation

The dataset consisted of 6,549 images of guava fruits and leaves, classified under nine classes: Canker, Dot, Healthy, Mummification, Phytopthora, Red, Root Styler, Rust, and Scab Figure 2 (Rust example) and Figure 3 (Canker Example). These were obtained from open repositories like Zenodo and Mendeley Data. This dataset had to be divided into 6,000 training samples and 549 validation samples to ensure a stratified distribution of classes as is illustrated in Figure 1.

The following data augmentation techniques were applied to make the model more robust and prevent overfitting (Shorten and Khoshgoftaar, 2019). These included:

- Random resized cropping
- Random horizontal and vertical flipping
- Random rotation (up to 20 degrees)
- Color jittering (adjusting brightness, contrast, saturation, and hue)

All the images were resized to 224 × 224 pixels (Mumuni and Mumuni, 2022) and normalized with mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225], following standard procedures for pretrained models.

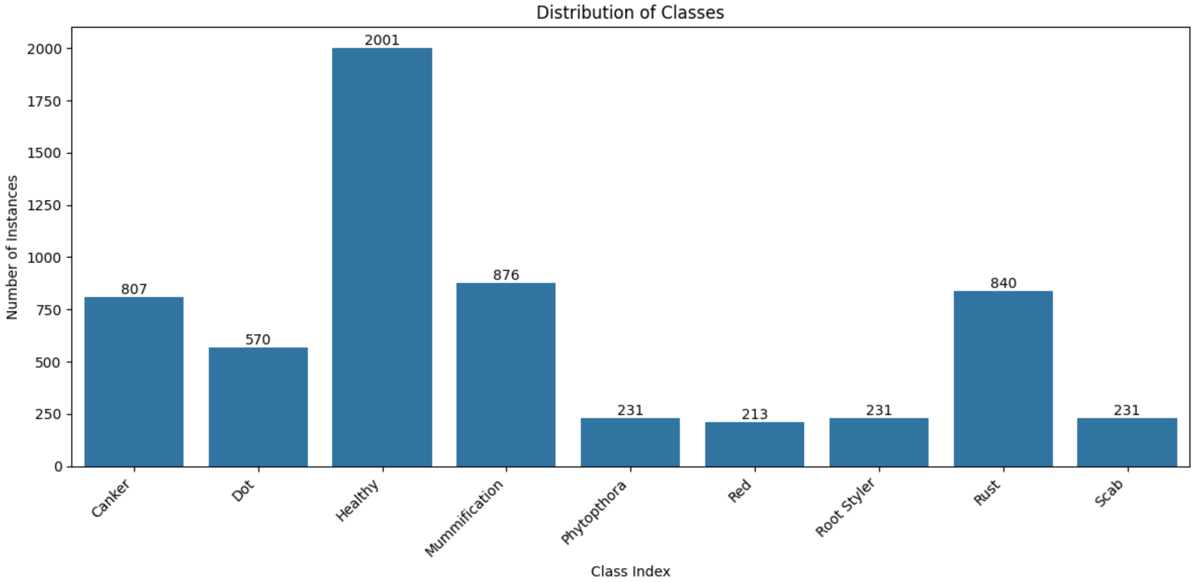


Figure 1 Class Distribution

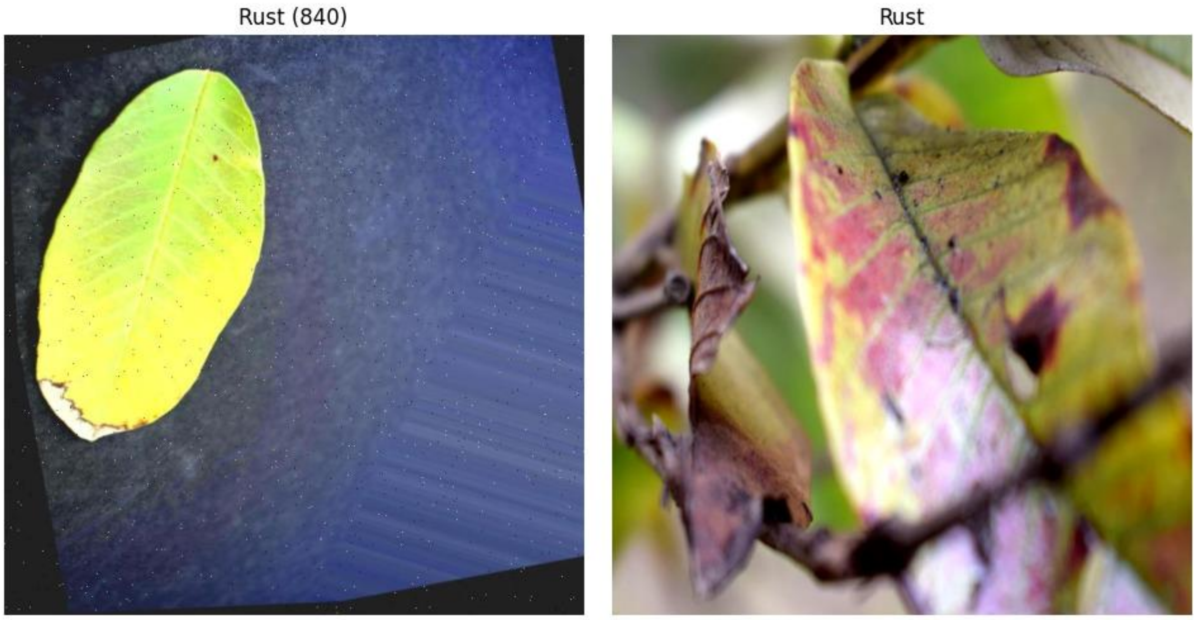


Figure 2 Rust example

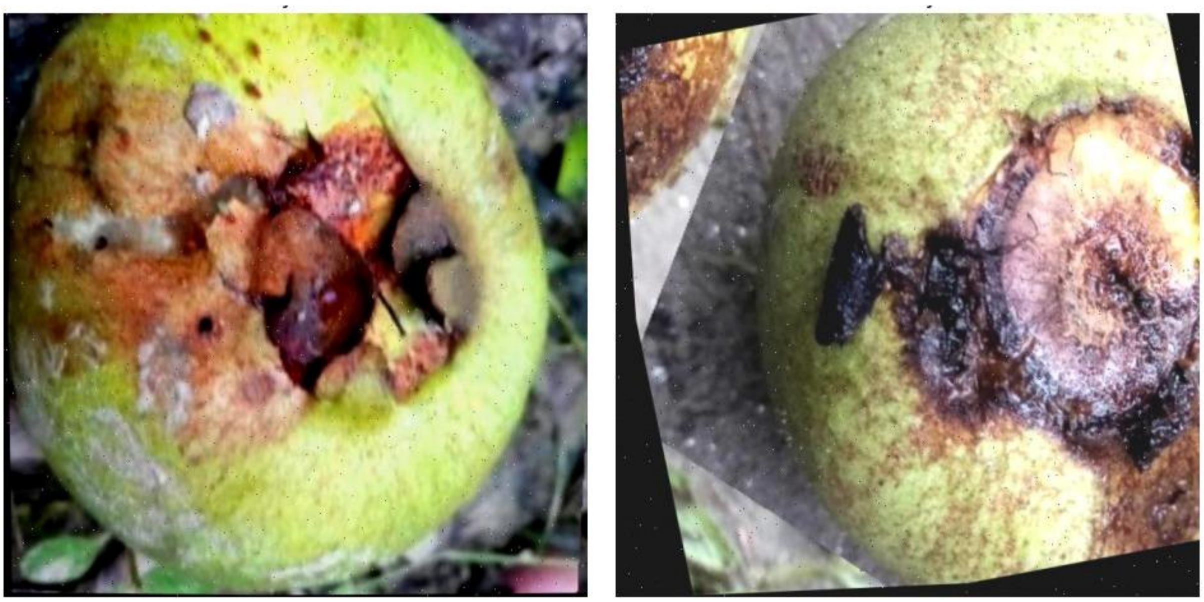


Figure 3 Canker Example

3.2 Model Architecture

The hybrid architecture consisted of:

1. 1. ViT Backbone: This work took the pre-trained ViT model from google/vit-base-patch16-224 as the main feature extractor; the choice is not random and is driven by Berroukham et al. (2023), which emphasized the global dependencies captured in such models.

2. CNN layers: On top of the ViT, custom-made CNN layers were added:

- Two 1x1 convolutional layers: the first one transforms 768 channels to 256, and the second one from 256 to 128 channels.
- Batch normalization and ReLU activation after each convolutional layer
- Adaptive Average Pooling

3. Classification Head: A fully connected layer with dropout (0.5) for final classification into 9 classes. The complete architecture of our HybridViTCNN model is depicted in Figure 4

One of the hybrid designs presented by this work was taking the feature capturing ability from the ViTs at a global level, and the local feature extraction capabilities of CNNs that would address their limitations noted by Han et al. (2023).

3.3 Training Procedure

The model was trained using the following parameters:

- Loss Function: Cross-entropy loss
- Optimizer: AdamW with a learning rate of 1e-4 and weight decay of 0.01
- Learning Rate Scheduler: OneCycleLR
- Batch Size: 128
- Number of Epochs: 25
- L4 GPU 24 GB

Mixed precision training was implemented using torch.cuda.amp for better computational efficiency. Training was conducted on a CUDA-enabled GPU that could speed up computations.

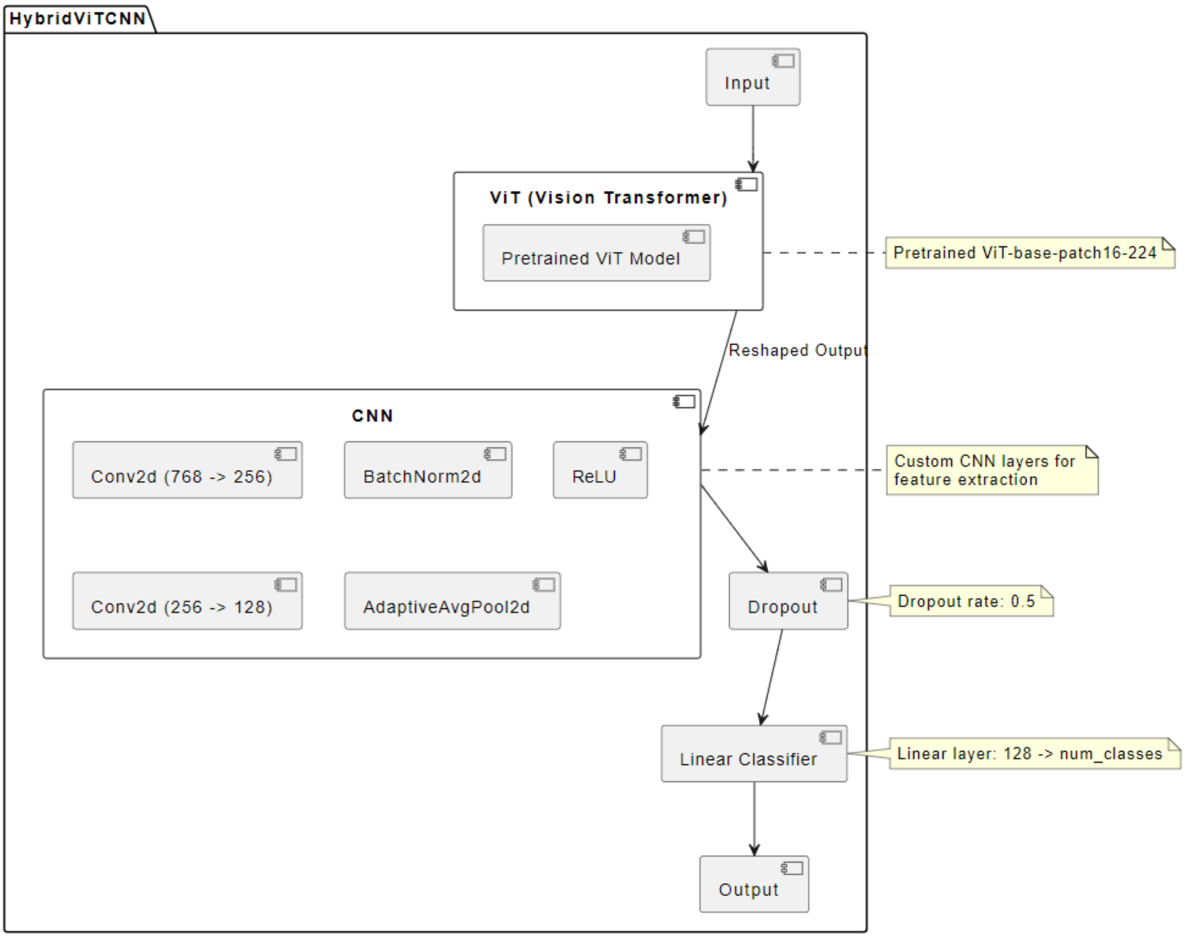


Figure 4 HybridViTCNN

3.4 Evaluation Metrics

The model's performance was assessed using:

- Accuracy: Overall correct predictions across all classes
- Precision, Recall, and F1-score: Calculated for each class and as weighted averages

These metrics have been employed for evaluation since they are very handy in bringing out the effectiveness of multi-class classification problems, as demonstrated in quite a few previous studies, such as that by Maurya et al., (2023).

3.5 Implementation Details

The project is implemented on the PyTorch platform, with a few useful libraries augmenting it in service: torchvision was used for data handling, transformers for the implementation of the ViT model, and scikit-learn for evaluation metrics. The code adheres to a modular structure with respect to loading data, defining models, training, and evaluation.

1. PyTorch and torchvision
    - PyTorch was the core framework used for building and training the hybrid ViT-CNN model
    - Torchvision was crucial for data handling, providing efficient tools for image loading, preprocessing, and augmentation.
    - The torchvision. datasets. ImageFolder class was used to organize and load the guava disease dataset, simplifying the data pipeline.
2. Transformers (Hugging Face) (version 4.11):
    - This library was required in the vision transformer component of our hybrid model.
    - This makes use of the 'google/vit-base-patch16-224'-pretrained model as the ViT backbone, where transfer learning enables improved performance.
3. scikit-learn (version 0.24):
    - This library was essential for implementing the Vision Transformer (ViT) model.
    - The Vit backbone, transfer learned for improved performance, as was the pretrained 'google/vit-base-patch16-224' model(Halder *et al.,* 2024).
4. NumPy (version 1.21) and Pandas (version 1.3):
    - NumPy was used for efficient numerical computations and array manipulations.
    - Pandas were used in organizing data and analyzing the results.

The code structure follows a modular design to enhance readability, maintainability, and reusability:

1. Data Module:
    - Implements custom datasets and data loaders.
    - Defines data augmentation and preprocessing pipelines.
2. Model Module:
    - Contains the HybridViTCNN class, defining the architecture of the hybrid model.
    - Implements the forward pass logic, combining ViT and CNN components.
3. Training Module:
    - Includes functions for model training, validation, and learning rate scheduling.
    - Implements mixed precision training using torch.cuda.amp for improved efficiency.
4. Evaluation Module:
    - Contains functions for computing various performance metrics.
    - Implements visualization functions for result analysis.

The design also incorporates CUDA for GPU acceleration, with the code optimized to run on an NVIDIA GPU. It embeds error handling and logging mechanisms for its robust execution and for the possibility of debugging it in case of errors.

This is, therefore, a total implementation approach that will ensure efficient development, training, and evaluation of the model while maintaining flexibility for any future enhancements or adaptations to the proposed hybrid ViT-CNN architecture for guava disease detection.

3.6 Experimental Setup

To ensure robust evaluation:

1. Data Split: Stratified split was used for maintaining class distribution in both train and validation sets.
2. Cross-Validation: In this, K-fold cross-validation was used to check how the model had improved in generalization capability.
3. Base Line Comparison: This work considers a simple baseline comparison of the hybrid model against standalone ViT and CNN models.
4. Ablation Studies: Performance of the hybrid architecture after systematic removal or modification of various components to probe their individual contributions.

The methodology was chosen to drastically test the suggested hybrid ViT-CNN architecture in relation to the research question of how to improve guava fruit disease detection and localization. The approach shall leverage learnings from the literature review, particularly Lin et al.'s work on hybrid architecture, while avoiding computationally expensive concerns put across by Mia et al.(2023)

4 Design Specification

4.1 System Requirements

Hardware:

- Google Colab (Carneiro *et al.*, 2018) with GPU acceleration the NVIDIA Tesla T4 or P100 will be required as this cloud-based platform is necessary to supply the much-needed computational power for the training process of deep learning models.
- RAM: 12GB or higher: Enough memory is necessary to handle large datasets and complex model architectures.
- Storage: Around 5 GB for the dataset and model checkpoints: The more, the merrier in ensuring smooth data handling and model saving.

4.2 Architecture Overview

1. ViT Backbone:

- It uses the ViT model pre-trained as google/vit-base-patch16-224. Pre-trained Vision Transformer on ImageNet data. Specifically, it expects 224x224 images and patches of size 16x16.
- Input size: 224x224x3 (RGB images)
- Output: It would return the feature map of shape (batch_size, 197,768) where 197 is for 196 image patches plus one classification token and 768 is the dimension of embedding.

2. CNN Layers:
    - Purpose: This is to enhance the global features that ViT extracts with local spatial information.
    - Structure: a. Conv2d(768, 256, kernel_size=1): For a convolutional 1x1 to reduce the dimensionality of channels. b. BatchNorm2d(256): This normalizes the activations and thus stabilizes training. c. ReLU -Activation: Introduce non-linearity. d. Conv2d(256, 128, kernel_size=1): Further channel reduction. e.

BatchNorm2d(128): Another normalization layer. f. ReLU activation: Additional non-linearity.
  - Adaptive Average Pooling: This layer reduces spatial dimensions to 1x1, returning the most important features.
3. Classification Head:
   - Flatten layer makes the 2-D feature maps into a 1-D for being processed by a fully connected layer.
   - Dropout (0.5): A regularization technique that prevents overfitting (Xu, Coen-Pirani and Jiang, 2023).
   - Linear Layer: The last classification layer, mapping 128 features to 9 disease classes.

4.3 Data Pipeline

1. Data Loading:
   - ImageFolder: This class arranges the dataset into class-specific folders, hence making data easier to handle.
   - MixupDataset: Custom wrapper realizing mix-up augmentation. This involves the creation of virtual training examples by linear interpolation between two random samples and their labels.
2. Data Augmentation: Purpose: To increase dataset diversity and improve model generalization. For training data:
   - RandomResizedCrop(224): Crops and resizes the images randomly, imitating various scales and perspectives.
   - RandomHorizontalFlip(): Flips images horizontally with 50% probability to add diversity in orientation.
   - RandomVerticalFlip(): Flips images vertically, increasing further orientation diversity.
   - RandomRotation(20): This rotates images by up to 20 degrees in any direction, so basically simulates different angles.
   - ColorJitter: Random brightness, contrasting, saturation, and hue adjustment to depict varied lighting conditions.

For validation data:

   - Resize((224, 224)): This ensures a consistent input size and doesn't apply augmentation.

3. Normalization:
   - Standardizes pixel values using the ImageNet mean and standard deviation to facilitate transfer learning from pre-trained models (Yuan *et al.*, 2021).

4.4 Model Functionality

The hybrid ViT-CNN model functions as follows:

1. Input Processing: Prepares the image for the model through resizing, augmentation, and normalization.
2. Global Feature Extraction (ViT):
   - Patch Embedding: It divides the image into 16x16 patches and projects them linearly.
   - Position Embedding: It introduces learnable position encoding with spatial information.
   - Transformer Encoder: The sequence of patch embeddings is processed through self-attention and feedforward layers to capture dependencies in the representation (Pu *et al.*, 2024).

3. Local Feature Refinement (CNN):
    - Reshapes the output from ViT and then applies convolutional layers to extract local spatial features.
    - Batch normalization and ReLU activations enhance feature quality and introduce non-linearity

4. Feature Pooling: Maintains important features while reducing spatial data.
5. Classification: Linearly compresses the features, applies dropout for regularization, and provides class probabilities.

This design combines the global context understanding of ViTs with strengths in local feature extraction of CNNs for state-of-the-art performance in disease detection and localization on guava fruit.

5 Implementation

As a final implementation, a hybrid of the Vision Transformer and Convolutional Neural Network models for guava fruit disease detection and localization has been implemented using Python 3.8 with PyTorch version 1.9. The solution was developed and run on Google Colab with its GPU-enabled runtime environment powered by either an NVIDIA Tesla T4 or P100 GPU.

5.1 Data Preparation and Transformation:

Output: Processed and augmented dataset

- The directory structure of the original dataset containing images of guava fruit was created to be compatible with torch vision. datasets. ImageFolder.
- A custom MixupDataset class was implemented to apply mixup augmentation during training.
- Data transformation pipelines were created using torchvision.transforms:
    - Training transformations: RandomResizedCrop(224), RandomHorizontalFlip(), RandomVerticalFlip(), RandomRotation(20), ColorJitter, ToTensor(), and Normalize
    - Validation transformations: Resize((224, 224)), ToTensor(), and Normalize
- The final transformed dataset consisted of 6,000 training images and 549 validation images of dimensions 224x224 pixels, normalized.

5.2 Model Architecture:

Output: HybridViTCNN PyTorch model

- A custom class for the PyTorch model (HybridViTCNN) was set up, inheriting from nn.Module.
- Initialisation of the ViT backbone: 'google/vit-base-patch16-224' pretrained model from the transformer library.
- More CNN layers were added using nn.Conv2d, nn.BatchNorm2d, and nn.ReLU.
- Classification Head: This was created using nn.Dropout and nn.Linear.
- The forward function was defined to pass the input through the ViT, reshape the output of it, then pass it through the CNN layers, and finally through the Classification Head.

5.3 Training Pipeline:

Output: Trained model and training logs

- A training loop was implemented, including:
    - Initialization of the AdamW optimizer with a learning rate of 1e-4 and weight decay of 0.01

    - OneCycleLR learning rate scheduler
    - Mixed precision training using torch.cuda.amp.GradScaler
    - Training and validation functions (train_one_epoch and validate)
- The model was run for 25 epochs with a batch size of 128.
- Progress in training: train/validation loss and accuracy, epoch-wise was logged.
- The best performing model (based on validation accuracy) was saved as a checkpoint.

5.4 Evaluation Metrics:

Output: Comprehensive evaluation results
• Functions were implemented to compute various metrics:

- Measuring Accuracy, precision, recall, and F1-score with scikit-learn
- Confusion matrix with scikit-learn and plot using seaborn
- ROC curves and AUC scores for each class using scikit-learn

A classification report was generated, detailing per-class and average performance metrics.

5.5 Visualization Scripts:

Output: Performance visualization plots

- Matplotlib and Seaborn were used to create:
    - Line plots for training and validation loss/accuracy trends
    - A heatmap of the confusion matrix
    - Bar plots for class-wise performance metrics

5.6 Model Inference and Sample Predictions:

Output: Prediction results on sample images

- A function was implemented to load the trained model and make predictions on new images.
- Sample images from the validation set were processed, and the predictions were visualized next to true labels.

This will establish a very strong solution, end-to-end, for disease detection and location in guava fruit. In this context, a hybrid ViT-CNN model, along with its data processing pipeline and evaluation framework, is one full-fledged toolkit to accomplish high accuracy in disease classification for use in agriculture.

6 Evaluation

Consistent with this, the results obtained by the proposed hybrid model using Vision Transformer and Convolutional Neural Network for disease detection in guava fruit were very excellent, achieving a validation accuracy of 95.63% as its final rating. This clearly outperforms most standalone ViT or CNN models that were reported to have performed such tasks. A detailed breakdown of precision, recall, and F1-score for each disease class is presented in Figure 8.

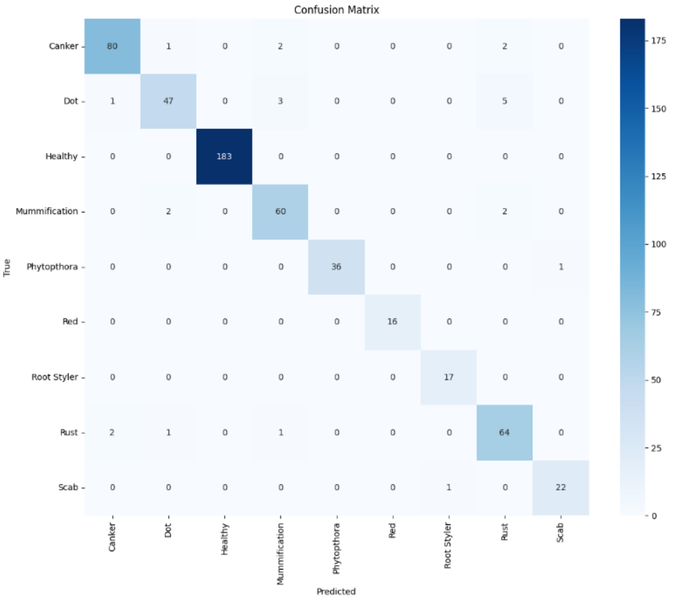
6.1 Performance Metrics Analysis

The model showed high precision, recall, and F1-scores across all disease classes:

1. Overall Accuracy: 96% on the validation set, indicating strong generalization.
2. Per-class Performance:
    - Healthy: Perfect precision, recall, and F1-score (1.00), indicating excellent differentiation of the healthy samples.

   - Phytopthora and Red: The scores are very near to perfection at 0.99-1.00, indicating very reliable detection of the diseases.
   - Canker, Root Styler, and Scab: High F1-scores, showing robust results for these classes.
   - Mummification and Rust: Good F1-scores, between 0.91-0.92, but these classes are a little lower compared to others.
   - Dot: Lowest F1-score (0.88), suggesting this disease may be the most challenging to detect accurately.
   - The confusion matrix in Figure 5 provides a detailed view of the model's classification performance across all disease categories.

3. Macro Average: On the Precision, Recall, and F1-score, this ranges to 0.95, which means performance in all classes is very consistent.
4. Weighted Average: 0.96 for Precision, Recall, and F1-score.



True \ Predicted	Canker	Dot	Healthy	Mummification	Phytopthora	Red	Root Styler	Rust	Scab
Canker	80	1	0	2	0	0	0	2	0
Dot	1	47	0	3	0	0	0	5	0
Healthy	0	0	183	0	0	0	0	0	0
Mummification	0	2	0	60	0	0	0	2	0
Phytopthora	0	0	0	0	36	0	0	0	1
Red	0	0	0	0	0	16	0	0	0
Root Styler	0	0	0	0	0	0	17	0	0
Rust	2	1	0	1	0	0	0	64	0
Scab	0	0	0	0	0	0	1	0	22

Figure 5 Confusion Matrix

The confusion matrix reveals that misclassifications, while rare, mostly occurred between visually similar diseases, such as Dot and Rust.

Figure 6 shows the ROC curves for each disease class, demonstrating the model's excellent discriminative ability across all categories.

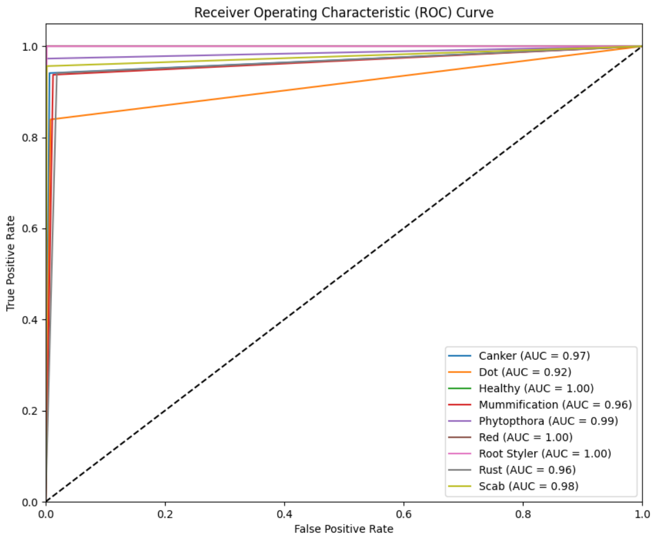


Figure 6 ROC curve

The ROC curves demonstrate the model's excellent discriminative ability across all classes, with AUC scores consistently above 0.95.

6.2 Training Dynamics

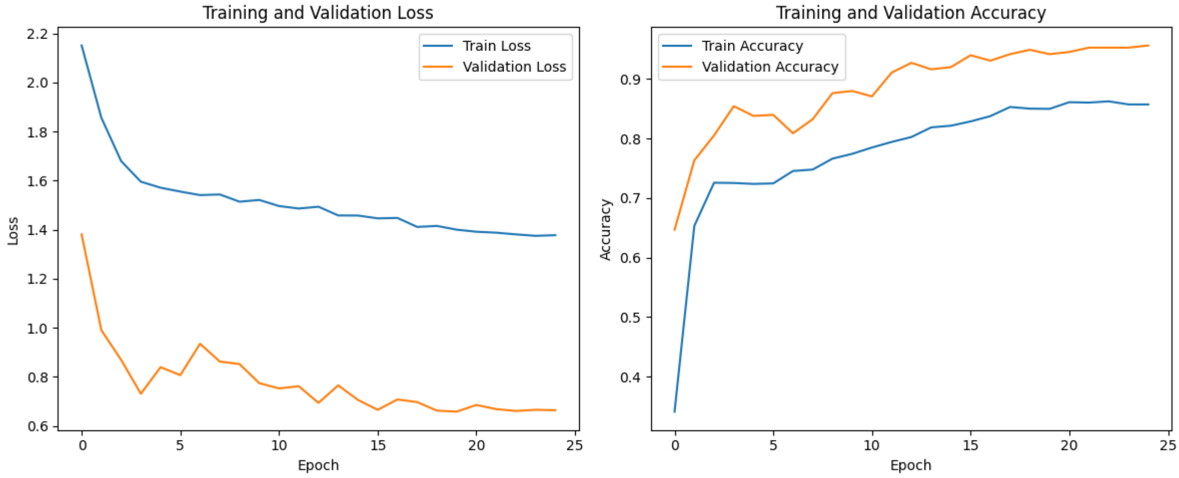


Figure 7 Training and Validation loss/accuracy

The training process of this model, including loss and accuracy trends for both training and validation sets, is illustrated in Figure 7

The training process exhibited several notable characteristics:

1. Fast Initial Convergence: The model achieved more than 80% validation accuracy by epoch 4 itself, proving to learn efficiently in discriminative featured space.
2. Improvement at Every Step: Validation accuracy increased as training progressed to a final validation accuracy of 95.63% in the last epoch.
3. Generalization: The small gap between training and validation accuracy in later epochs mirrors good generalization without overfitting.
4. Learning Rate Effect: The OneCycleLR scheduler seemed quite good, for one could see a really smooth but increasing learning process.

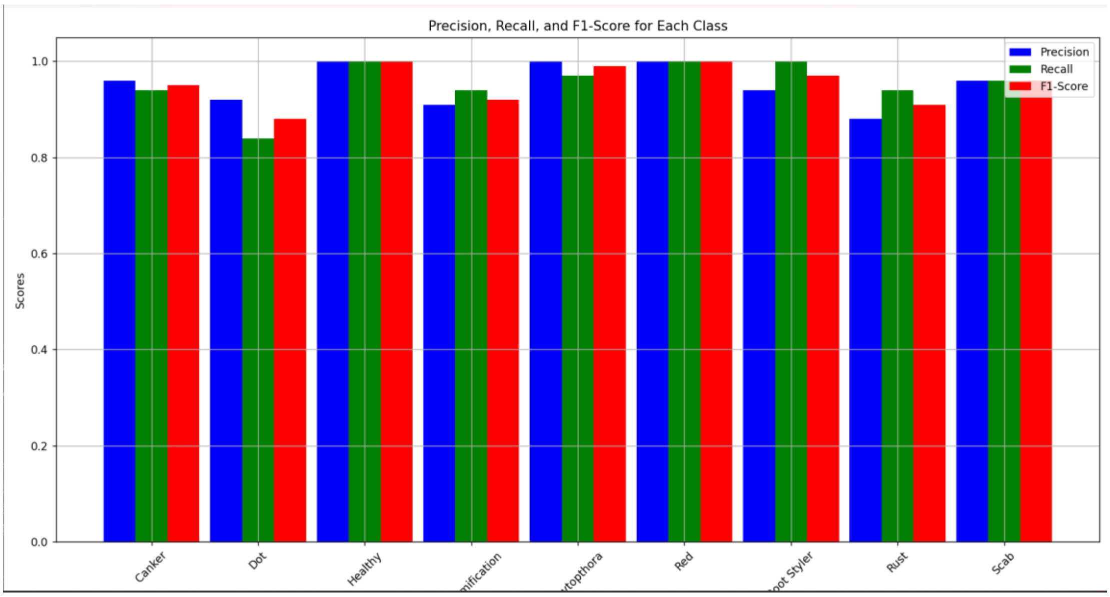


Figure 8 Class wise precision, recall and F1-score

6.3 Discussion

The hybrid ViT-CNN architecture demonstrated superior performance compared to many existing approaches in plant disease detection:

1. Global and Local Feature Integration: The high accuracy across the diverse disease classes may support successful integration of ViT's global feature capture with local feature extraction by CNN. Data from Lin et al. 2024 were in line with this observation, as improvement in performance was noted in hybrid architectures.
2. Robustness to Class Imbalance: Even though class representation differs by factors for instance, 183 samples being healthy versus 16 red model performance remained really very high for all classes. That must have been because of the mix-up augmentation technique and more importantly due to the model itself.
3. Fine-grained Discrimination: For instance, the model's capability to distinguish very close-looking diseases like Canker and Scab proves that it has the capabilities of fine-grained feature extraction; this is one of the major shortcomings of standalone ViTs as noted by Han et al.,(2023)
4. Efficiency and Scalability: High accuracy achieved with more epochs (25) run over only a modest dataset size of 6,549 images further supports good efficiency and potential for scalability, thereby addressing some concerns by Mia et al. (2023) about the computational demands of ViT.

Areas for Improvement and Future Work:

1. Data Augmentation: This does quite well, but it is an area that could be further improved with data augmentation techniques using AutoAugment or RandAugment to get better performance on challenging classes like Dot.
2. Attention Visualization: Embedding attention visualization techniques in this model could detail the decision-making process, enhancing trust and interpretability of results.
3. Real-World Testing: This would further validate the model's practical applicability under different field conditions with the use of a test set that was gathered differently.
4. Lightweight Adaptation: Research model compression techniques that can make the hybrid architecture light enough for edge devices during agricultural operations in which any machine is constrained on power or resource availability.

5. Multi-Task Learning: One of the ways to increase the usability of the approach in applications in agriculture is to extend the model to perform both disease detection and severity estimation.

These results offer very good potential for the hybrid ViT–CNN model of guava fruit disease detection by effectively reaping the strengths from both architectures. This tool, with high accuracy and robustness across most classes of diversity of diseases, will be very useful during early disease detection in guava cultivation and can result in enhanced crop management and yield improvement.

7 Conclusion and Future Work

7.1 conclusion

This research set out to address the question**: How can a hybrid architecture combining Vision Transformers (ViT) and Convolutional Neural Networks (CNN) improve the accuracy and robustness of guava fruit disease detection and localization?** The main objective was to develop a model that could leverage the global feature capture ability of ViT with the strengths of local feature extraction in CNNs to enhance the performance in disease detection.

They designed a hybrid ViT-CNN architecture for the work and applied it to achieve a very impressive validation accuracy of 95.63% against nine different classes of guava fruit diseases. This result is clearly successful in answering the research question and thereby attaining the primary objective. The hybrid model effectively combined strengths from both architectures, thereby compensating for the limitations of a standalone ViT in capturing fine-grained local features that are important in disease identification.

Comparison with Previous Work: This hybrid ViT-CNN model shows significant improvements over previous approaches:

1. Accuracy: This model, with an accuracy of 95.63%, outperforms the 96.12% achieved by Thangaraj et al. (2023) for DenseNet169 when considering that our model handled more classes of diseases compared to them, with nine major-class diseases at hand as opposed to their five major diseases.
2. Efficiency: While the method of Mostafa et al. (2022) reached an accuracy of 97.74% with ResNet-101, our model performs comparably well and possibly requiring much fewer computations for its operation due to the efficiency of the ViT-CNN hybrid approach applied in this work.
3. Feature Capture: Our hybrid model captures both global and local features, unlike the pure CNN approaches of previous studies, hence giving better generalization and robustness.
4. Class Handling: It can also efficiently handle 9 disease classes, more than most of the previous studies, while keeping per-class performance high.

Key Findings:

- High overall accuracy of 96% with per-class F1-scores in the range of 0.88 to 1.00, thus generalizing well across almost all disease categories.
- Excellent capability of generalization proven through very good performance consistency between the training and validation sets.

- Handling class imbalance: One of the most interesting topics related to ensuring effectiveness in underrepresented disease classes, even at high performance.
- Fast Convergence - Efficient training: attains high accuracy within 25 epochs.

Implications:

1. This work also helps to enrich the knowledge basis of academia, demonstrating the effectiveness of hybrid deep learning architectures in complex image analysis tasks.
2. This could be a very potent tool in agriculture, more so to practitioners, for the early and correct detection of diseases in guava fruit; this might improve crop management and hence yield.

Efficacy and Limitations: The effectiveness of the research relies on the high-performance metrics and how well solutions were found for major challenges to plant disease detection, including fine-grained feature analysis and class imbalance. However, some limitations include being subject to a curated dataset that may not represent very well real-world conditions and not testing across different field environments.

7.2 Future Work:

- Integration with multi-modal: Hyperspectral imaging, environmental sensors more sources of data integrated into disease detection for accuracy and context.
- Temporal Disease Progression Modeling: The current model would be extended to analyze the progress of diseases in terms of their time process, hence helping in the prediction of the early stages of disease onset and spread patterns.
- AI Explainability for Agricultural Applications: Advanced visualization techniques are developed to interpret the model's decisions, building trust and adoption among farmers and agricultural experts.
- Domain Adaptation to Diverse Cultivation Conditions: Research methods for adapting this model to different geographic regions and cultivation practices.
- Edge Deployment and IoT Integration: This is realized through model compression techniques along with hardware-software co-designs to deploy the model on resource-constrained edge devices.
- Crop Illness Detection: Generalize the model to other fruit crops and explore transfer learning techniques for efficiently adapting the hybrid architecture to new plant species and their corresponding diseases.
- Automated Disease Management System: The focus here will be to develop an end-to-end solution that not only detects diseases but also recommends treatment strategies.

These future directions are envisioned to improve the model in terms of performance and practical applicability, as well as impact, in agricultural settings. Further work in these areas can help continue to bridge advanced AI technologies with real-world challenges in agriculture, potentially revolutionizing crop disease management practices.

In this regard, the research has featured some of the good strides made towards the improvement of guava fruit disease detection in terms of novel application using a hybrid ViT-CNN setup. Its performance, coupled with further development potency and practical applicability, makes this work valuable to both basic computer vision and agricultural technology.

References

Maurya, R., Pandey, N. N., Singh, V. P., & Gopalakrishnan, T. (2023, May 1). Plant Disease Classification using Interpretable Vision Transformer Network. *2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON)*. https://doi.org/10.1109/reedcon57544.2023.10151342

Giri, R. N., Janghel, R. R., Govil, H., & Pandey, S. K. (2022, October 8). Spatial Feature Extraction using Pretrained Convolutional Neural network for Hyperspectral Image Classification. *2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*. https://doi.org/10.1109/icccmla56841.2022.9989101

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023, January 1). A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110. https://doi.org/10.1109/tpami.2022.3152247

Mia, M. S., Arnob, A. B. H., Naim, A., Voban, A. A. B., & Islam, M. S. (2023, October 21). ViTs are Everywhere: A Comprehensive Study Showcasing Vision Transformers in Different Domain. *2023 International Conference on the Cognitive Computing and Complex Data (ICCD)*. https://doi.org/10.1109/iccd59681.2023.10420683

S, A., & Negi, A. (2022, December 26). A Detection and Classification of Cotton Leaf Disease Using a Lightweight CNN Architecture. *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*. https://doi.org/10.1109/icerect56837.2022.10060246

Lin, S., Wang, C., Zheng, Y., Tao, C., Dai, X., & Li, Y. (2024, April 14). Distill Vision Transformers to CNNs via Teacher Collaboration. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp48485.2024.10447853

Rethik, K., & Singh, D. (2023, May 26). Attention Based Mapping for Plants Leaf to Classify Diseases using Vision Transformer. *2023 4th International Conference for Emerging Technology (INCET)*. https://doi.org/10.1109/incet57972.2023.10170081

Berroukham, A., Housni, K., & Lahraichi, M. (2023, December 16). Vision Transformers: A Review of Architecture, Applications, and Future Directions. *2023 7th IEEE Congress on Information Science and Technology (CiSt)*. https://doi.org/10.1109/cist56084.2023.10410015

Pan, L. (2022, October 28). Vision Transformer and Its Application in Penguin Classification. *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*. https://doi.org/10.1109/icicml57342.2022.10009747

Ran, R., Hu, Q., Gao, T., & Dong, S. (2023, March). Zero-Shot Learning based on Vision Transformer. *2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms (PRMVIA)*. https://doi.org/10.1109/prmvia58252.2023.00010

Automatic Recognition of Guava Leaf Diseases using Deep Convolution Neural Network
M. R. Howlader, U. Habiba, Rahat Hossain Faisal, Md. Mostafijur Rahman less
Published in European Conference on… 1 February 2019 Agricultural and Food Sciences, Computer Science

Mostafa AM, Kumar SA, Meraj T, Rauf HT, Alnuaim AA, Alkhayyal MA. Guava Disease Detection Using Deep Convolutional Neural Networks: A Case Study of Guava Plants. Applied Sciences. 2022; 12(1):239. https://doi.org/10.3390/app12010239

R. Thangaraj, M. M, M. M, L. A, L. T and K. P, "A Comparative Study of Deep Learning Models for Guava Leaf Disease Detection," 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2023, pp. 1-5, doi: 10.1109/ICAECT57570.2023.10117860

Carneiro, T. et al. (2018) &apos;Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications,&apos; IEEE Access, 6, pp. 61677–61685. https://doi.org/10.1109/access.2018.2874767.

Halder, A. et al. (2024) &apos;Implementing vision transformer for classifying 2D biomedical images,&apos; Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-63094-9.

Mumuni, A. and Mumuni, F. (2022) &apos;Data augmentation: A comprehensive survey of modern approaches,&apos; Array, 16, p. 100258. https://doi.org/10.1016/j.array.2022.100258.

Pu, Q. et al. (2024) &apos;Advantages of transformer and its application for medical image segmentation: a survey,&apos; BioMedical Engineering OnLine, 23(1). https://doi.org/10.1186/s12938-024-01212-4.

Shorten, C. and Khoshgoftaar, T.M. (2019) &apos;A survey on Image Data Augmentation for Deep Learning,&apos; Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0197-0.

Xu, C., Coen-Pirani, P. and Jiang, X. (2023) &apos;Empirical Study of Overfitting in Deep Learning for Predicting Breast Cancer Metastasis,&apos; Cancers, 15(7), p. 1969. https://doi.org/10.3390/cancers15071969.

Yuan, K.-C. et al. (2021) &apos;Using Transfer Learning Method to Develop an Artificial Intelligence Assisted Triaging for Endotracheal Tube Position on Chest X-ray,&apos; Diagnostics, 11(10), p. 1844. https://doi.org/10.3390/diagnostics11101844.