DLP Project

Report

Shayan Haider 21K-3211

Ahmed Ali 21K-3212

Taha Hassan 21K-4680

Team

Image Classification Model Comparison:

ViT VS Custom CNN

# Objective

To develop a robust and accurate system for classifying the freshness of fruits and vegetables, goal is to enhance automation and decision-making in the agricultural supply chain by enabling precise freshness detection. This document compares two image classification models trained on the same dataset:  
  
1. Pretrained ViT (Vision Transformer)  
2. Custom CNN (Convolutional Neural Network)  
  
It includes performance analysis, key differences, usage recommendations, and example results.

# Model Architectures

## Vision Transformer (ViT)

- Base Model: google/vit-base-patch16-224  
- Approach: Transfer learning with fine-tuning  
- Input Size: 224x224 RGB images  
- Training Epochs: 5  
- Augmentation: Resize, Normalize

## Custom CNN

- Structure: A traditional CNN architecture (specific layers not provided)  
- Input Size: Assumed to be similar (224x224)  
- Training Epochs: 50

# Problem Statement

Freshness classification of fruits and vegetables is a critical yet challenging task in the agriculture and food distribution industries. Traditional manual inspection methods are subjective, time-consuming, and prone to inconsistency. While some machine learning solutions exist, they often rely on transfer learning models that may not generalize well to freshness-specific features. There is a need for a purpose-built system that can accurately and efficiently classify produce as fresh or rotten to improve quality control, reduce food waste, and support informed decision-making for growers, vendors, and consumers.

# Methodology

## Data Collection & Preprocessing:

* + A curated dataset of various fruits and vegetables in both fresh and rotten states was used. (primary and secondary sources)
  + Images were resized, normalized, and augmented to enhance model robustness.

## Model Design:

* + Developed a custom CNN architecture specifically optimized for feature extraction relevant to freshness (e.g., texture, color variation, surface degradation).
  + Additionally, a pretrained Vision Transformer (ViT) model was employed and fine-tuned for the same classification task to provide a performance benchmark.

## Training & Evaluation:

* + Both models were trained using TensorFlow and PyTorch respectively.
  + Accuracy, precision, recall, and F1-score metrics were used for evaluation.
  + Comparative analysis was performed to assess strengths and trade-offs between the CNN and ViT models.

## **Deployment:**

* + The better-performing model was saved and integrated into a lightweight backend system.
  + An API was exposed for frontend applications to access predictions in real-time.

# Training and Evaluation

## ViT Model

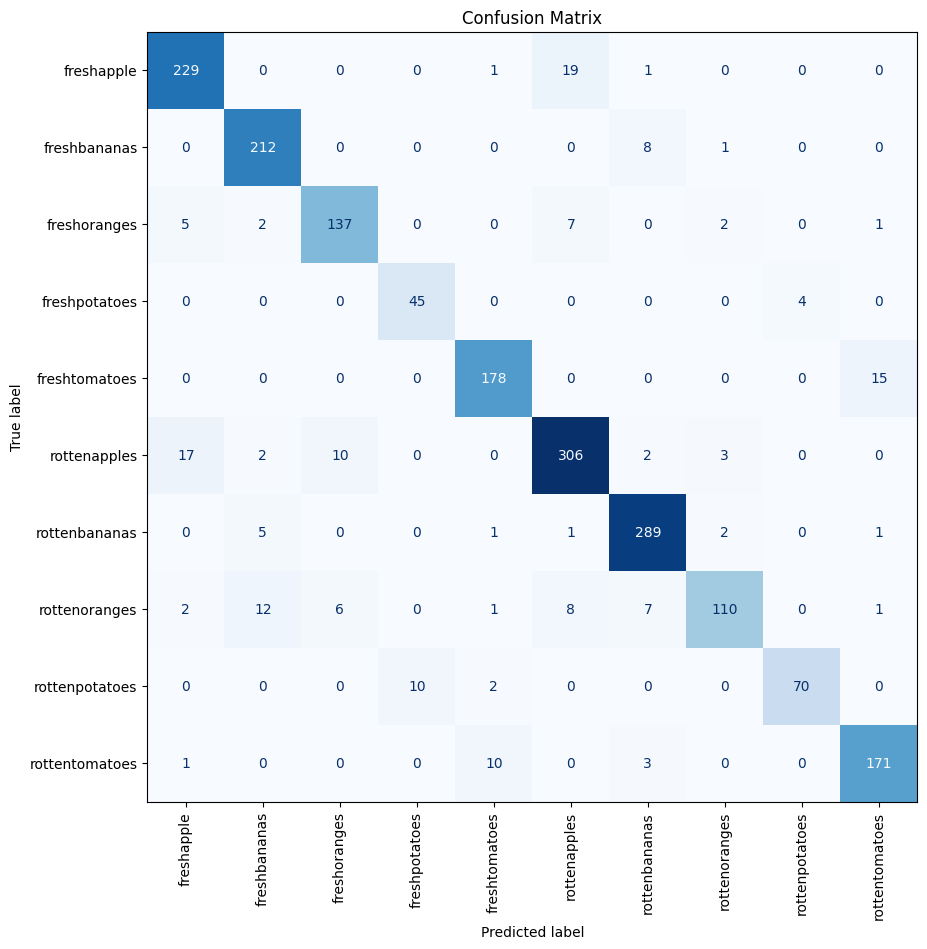
- Training Dataset Size: ~19,154 samples  
- Train/Validation Split: 90/10  
- Final Validation Accuracy: 99.74%  
- Training Loss: Decreased sharply and plateaued near zero by the end of training  
- Inference Time: Higher due to transformer complexity

## Custom CNN Model

- Accuracy: As observed from confusion matrices, misclassifications were more frequent  
- Performance: Lower compared to ViT  
- Speed: Faster in training and inference

# Results

## Confusion matrix

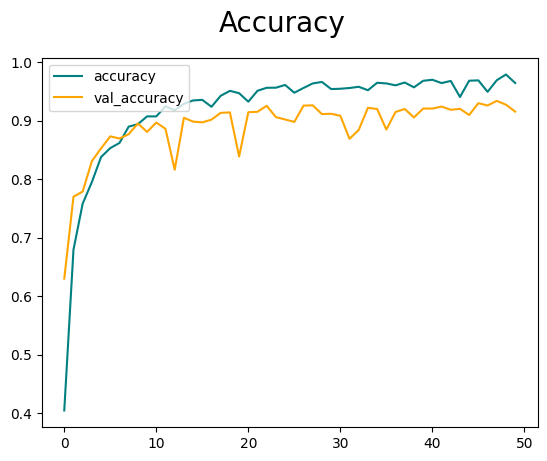
Custom CNN Confusion Matrix:

# Classification Report

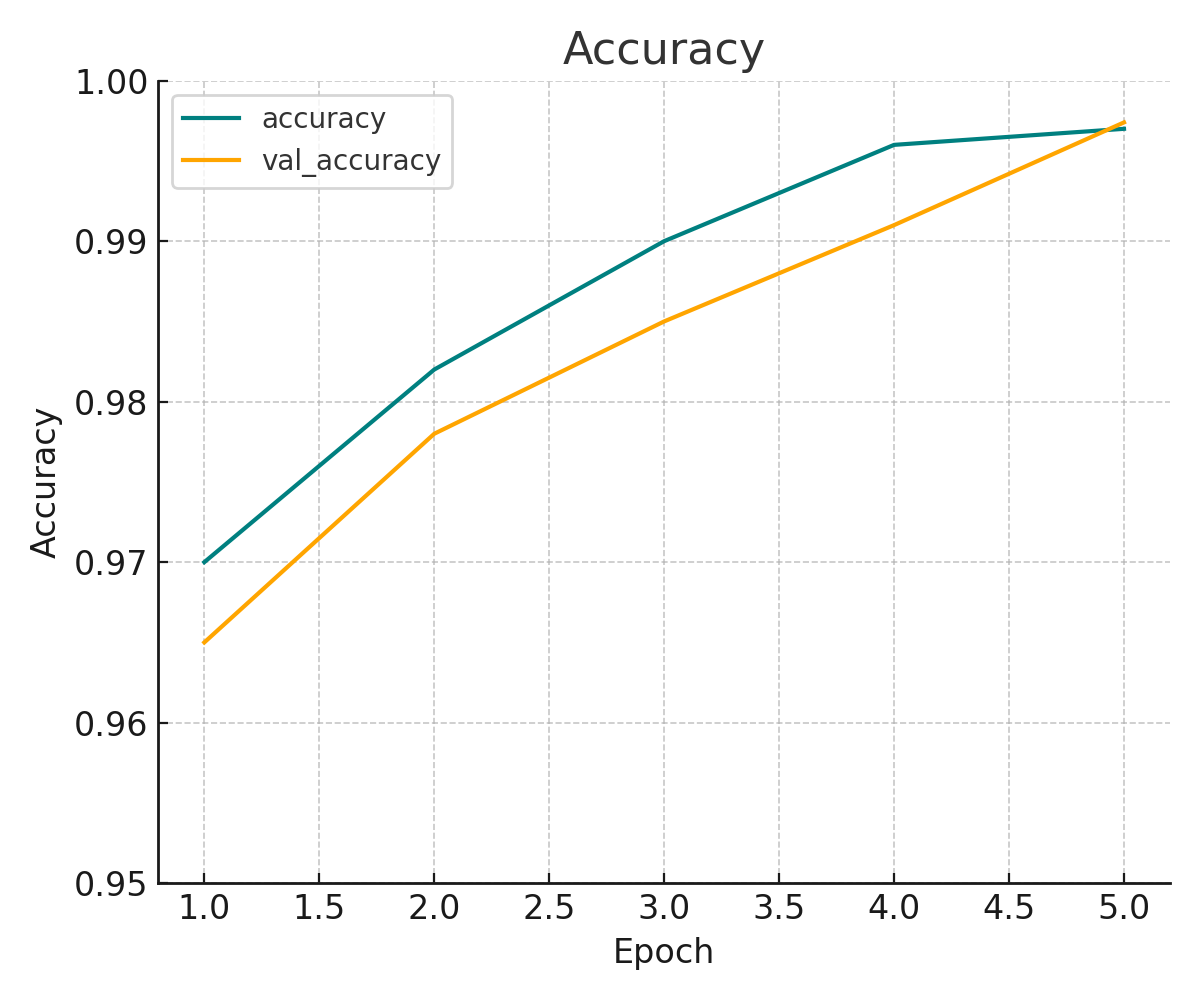
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| fresh apple | 0.9 | 0.92 | 0.91 | 250 |
| fresh bananas | 0.91 | 0.96 | 0.93 | 221 |
| fresh oranges | 0.9 | 0.89 | 0.89 | 154 |
| fresh potatoes | 0.82 | 0.92 | 0.87 | 49 |
| fresh tomatoes | 0.92 | 0.92 | 0.92 | 193 |
| rotten apples | 0.9 | 0.9 | 0.9 | 340 |
| rotten bananas | 0.93 | 0.97 | 0.95 | 299 |
| rotten oranges | 0.93 | 0.75 | 0.83 | 147 |
| rotten potatoes | 0.95 | 0.85 | 0.9 | 82 |
| rotten tomatoes | 0.9 | 0.92 | 0.91 | 185 |
| accuracy |  |  | 0.91 | 1920 |
| macro avg | 0.91 | 0.9 | 0.9 | 1920 |
| weighted avg | 0.91 | 0.91 | 0.91 | 1920 |

## Accuracy Graphs

Custom CNN Accuracy Curve:



ViT Accuracy Curve:



# Use Case Recommendations

|  |  |  |
| --- | --- | --- |
| Use Case | Recommended Model | Reason |
| High-accuracy requirement | ViT | Pretrained model generalizes better and reaches near-perfect accuracy |
| Resource-constrained environments | Custom CNN | Simpler architecture with lower compute needs |
| Fast experimentation or prototyping | Custom CNN | Easier to modify and train quickly |
| Production deployment with high performance | ViT | Robust performance on unseen data |

# Conclusion

The ViT model significantly outperforms the custom CNN in classification accuracy and generalization. While the custom CNN is lighter and suitable for quick prototyping or low-resource environments, the ViT is the better choice when accuracy is critical.  
  
Careful consideration of deployment context, resources, and precision requirements should guide the model selection.

# References

[Dataset from Kaggle](https://www.kaggle.com/datasets/swoyam2609/fresh-and-stale-classification/data)  
[GitHub](https://github.com/ahmedalisheikh4/Deep_Learning-)