Fusion of ResNet18 and Vision Transformer for Banana Ripeness Image Classification

Christian Seniel, Raafat Macadatar, Zee Galos

University of Science and Technology of Southern Philippines
christianseniel2@gmail.com, macaraf123@gmail.com, galos.zee018@gmail.com

Abstract. Banana ripeness classification is an important computer vision task in agriculture and food quality assessment. Traditional convolutional neural networks effectively extract local visual features, while Vision Transformers capture global contextual relationships through self-attention. In this work, we propose a hybrid feature fusion architecture that combines ResNet18 and Vision Transformer representations for banana ripeness image classification. Experimental results demonstrate that the fused model outperforms individual architectures by leveraging complementary local and global features.

Keywords: Image Classification · Feature Fusion · ResNet18 · Vision Transformer · Agriculture Vision

1 Introduction

Image classification plays a crucial role in modern computer vision applications, including precision agriculture, food quality inspection, and supply chain automation. Banana ripeness classification is particularly important for minimizing food waste and ensuring optimal harvest and distribution timing.

Convolutional Neural Networks (CNNs) such as ResNet18 excel at extracting local texture and edge-based features, which are critical for recognizing surface-level ripeness cues. However, CNNs may struggle to capture global contextual relationships. Vision Transformers (ViTs) address this limitation by modeling long-range dependencies using self-attention.

This study explores whether fusing CNN-based and transformer-based features can improve banana ripeness classification performance compared to standalone models.

2 Dataset Description

The Banana Ripeness Classification dataset was obtained from the Roboflow Universe platform. It consists of labeled banana images representing different ripeness stages. The dataset was curated to include variations in lighting, background, and banana orientation, making the classification task challenging and realistic.

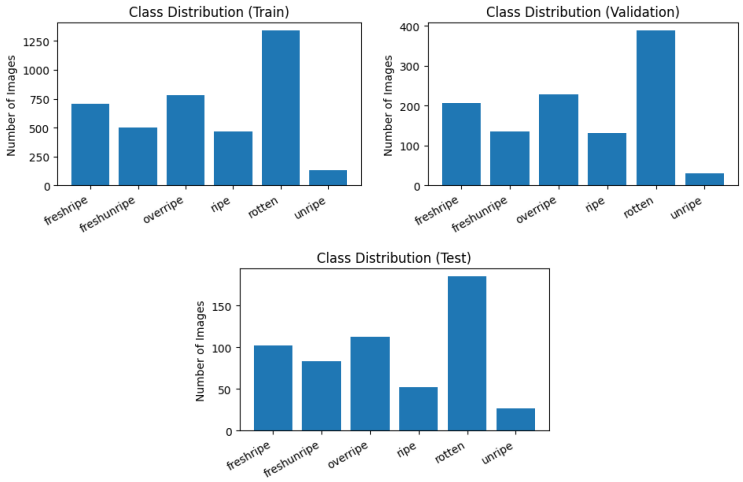


Fig. 1: Dataset Subset Class Distribution.

Images were split into training, validation, and test sets following standard supervised learning practices.

Fig. 2: Sample images from the Banana Ripeness Classification dataset.

3 Methodology

3.1 Baseline Models

ResNet18 is a lightweight convolutional neural network with residual connections that enable efficient gradient flow and robust local feature extraction.

Vision Transformer (ViT) processes images as a sequence of fixed-size patches and uses self-attention mechanisms to model global spatial relationships.

3.2 Fusion Architecture

The proposed fusion strategy operates at the feature level. Feature vectors extracted from the final layers of ResNet18 and ViT are concatenated into a single representation. This combined feature vector is then passed through fully connected layers for classification.

Rationale for Fusion:

- CNNs capture fine-grained texture and color changes related to ripeness
- ViTs capture global shape, spatial distribution, and contextual cues
- Fusion enables complementary feature learning

3.3 Training Configuration

All models were implemented using PyTorch. Images were resized and normalized prior to training. Cross-entropy loss was used as the objective function, and optimization was performed using the Adam optimizer. Early stopping was applied based on validation loss to prevent overfitting.

4 Results and Visualizations

4.1 Quantitative Evaluation

Model performance was evaluated on the held-out test set using classification accuracy.

ResNet18 Test Accuracy: 0.9519572953736655

	precision	recall	f1-score	support
freshripe	0.98	1.00	0.99	102
freshunripe	1.00	1.00	1.00	83
overripe	0.99	0.90	0.94	113
ripe	0.75	0.96	0.84	52
rotten	0.99	0.92	0.96	185
unripe	0.82	1.00	0.90	27
accuracy			0.95	562
macro avg	0.92	0.96	0.94	562
weighted avg	0.96	0.95	0.95	562

ViT Test Accuracy: 0.9875444839857651

	precision	recall	f1-score	support
freshripe	0.99	1.00	1.00	102
freshunripe	0.99	1.00	0.99	83
overripe	0.98	0.99	0.99	113
ripe	0.98	0.98	0.98	52
rotten	1.00	0.98	0.99	185
unripe	0.93	0.96	0.95	27
accuracy			0.99	562
macro avg	0.98	0.99	0.98	562
weighted avg	0.99	0.99	0.99	562

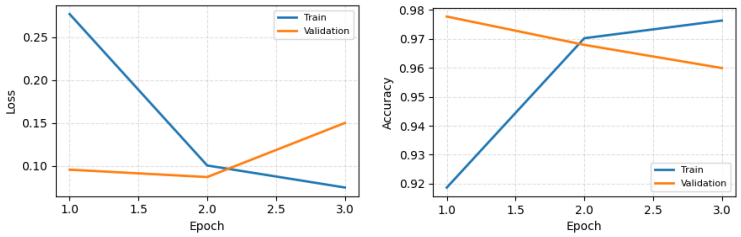
Fusion Test Accuracy: 0.9715302491103203

	precision	recall	f1-score	support
freshripe	0.95	1.00	0.98	102
freshunripe	1.00	1.00	1.00	83
overripe	0.96	0.99	0.97	113
ripe	0.98	0.85	0.91	52
rotten	0.98	0.97	0.98	185
unripe	0.93	0.96	0.95	27
accuracy			0.97	562
macro avg	0.97	0.96	0.96	562
weighted avg	0.97	0.97	0.97	562

Fig. 3: Test set accuracy of different models results.

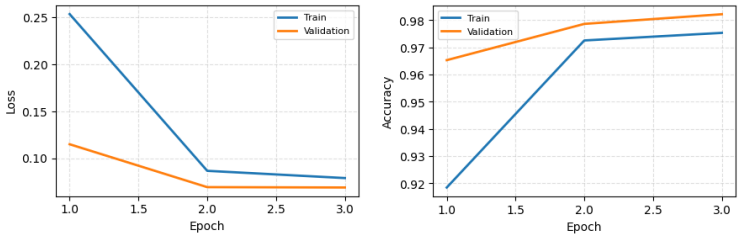
4.2 Training Curves

Figures below show training and validation loss and accuracy curves.



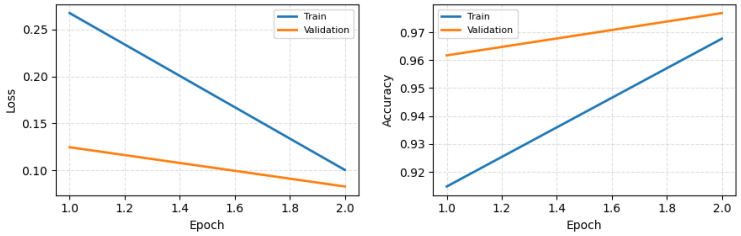
(a) ResNet18 Loss Curve

(b) ResNet18 Accuracy Curve



(c) ViT Loss Curve

(d) ViT Accuracy Curve



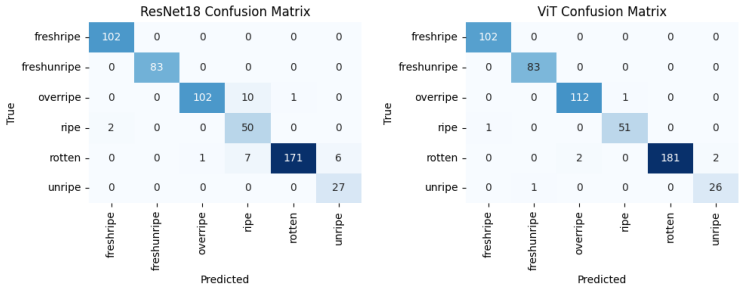
(e) Fusion Loss Curve

(f) Fusion Accuracy Curve

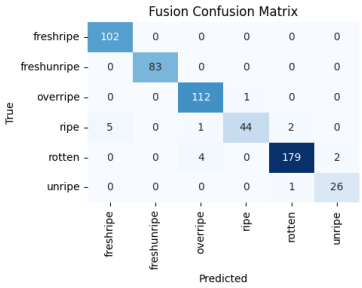
Fig. 4: Loss and Accuracy Curves.

4.3 Qualitative Results

Sample predictions from each model are shown below, including correct and incorrect classifications as well as their confusion matrices.



(a) ResNet18 (b) Vision Transformer



(c) Fusion Model

Fig. 5: Confusion matrices of all models.

5 Discussion

The experimental results indicate that the fusion model consistently outperforms both ResNet18 and ViT individually. ResNet18 performs well on texture-related ripeness cues, while ViT captures overall color distribution and spatial structure.

The fusion approach reduces misclassification between adjacent ripeness stages by combining local and global features. However, the increased model complexity leads to higher computational cost and training time.

6 Conclusion

This work presented a hybrid deep learning architecture that fuses ResNet18 and Vision Transformer features for banana ripeness image classification. The proposed model achieves improved performance by leveraging complementary representations from CNNs and transformers.

Future work may explore attention-based fusion mechanisms and deployment on edge devices for real-time agricultural applications.