Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Banana Ripeness Stage Identification Using Transformers

Rudresh Arabekar Rudresh Arabekar
Sahil Goes Sahil Goes
Anant Sawant Anant Sawant
22co51@aitdgoa.edu.in

Vollny Coelho Vollny Coelho
Kedar Sawant Kedar Sawant5
Tanvi Dharwadkar Tanvi Dharwadkar

Research Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10030843/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Banana Ripeness Stage Identification Using Transformers

Rudresh Arabekar[1], Sahil Goes[2], Anant Sawant[3 *], Vollny Coelho[4], Kedar Sawant[5], Tanvi Dharwadkar[6]

Agnel Institute of Technology and Design, Mapusa- Bardez, Goa, India, 403507
(Corresponding author email: 22co51@aitdgoa.edu.in)*

Abstract:

The ripeness level of bananas significantly influences their shelf life, market value, and suitability for consumption. Traditional ripeness assessment methods are manual, subjective, and often inconsistent, leading to increased post-harvest losses and inefficient supply chain management. This study proposes an automated banana ripeness classification system using deep learning models, with a focus on Vision Transformers (ViTs), to analyze high-resolution images of Moira (Myndoli) bananas. The system categorizes bananas into five ripeness stages: Unripe, Midripe, Ripe, Overripe, and Rotten, based on visual features such as color, texture, and surface patterns.A dataset of 4,500 labeled images (900 per class) was utilized, with preprocessing steps including background removal, resizing to 224×224, and normalization. Controlled data augmentation techniques—such as horizontal flipping, rotation, and mild brightness and contrast adjustments—were applied to improve generalization while preserving color-sensitive features. Five deep learning architectures were evaluated: ViT-Base, Swin Transformer Tiny, CvT, ResNet50, and MobileNetV3. Model performance was assessed using accuracy, precision, recall, F1-score, and confusion matrix analysis under a stratified 5-fold cross-validation framework.Among all models, the Swin Transformer Tiny achieved the best performance, with an average cross-validation accuracy of 87.39% and a final test accuracy of 86.78%, demonstrating both high accuracy and stability. The selected model was further enhanced with Explainable AI (XAI) techniques, utilizing EigenCAM as the primary method and Grad-CAM for supplementary analysis, to generate visual explanations of predictions.The proposed system provides a reliable and interpretable solution for automated fruit grading, offering practical benefits for farmers, vendors, and supply chain stakeholders.

Keywords: Banana ripeness classification; Swin Transformer; Vision Transformer; Explainable AI; Grad-CAM; EigenCAM; Deep learning; Transfer learning; Stratified cross-validation; Precision agriculture

1. Introduction

Bananas rank among the most economically significant tropical fruits worldwide, contributing substantially to global food security and international agricultural trade. According to the Food and Agriculture Organization of the United Nations, global banana production has exceeded 120 million tonnes annually in recent years, making it the fourth largest agricultural commodity by production volume [1]. Post-harvest losses attributed to inadequate ripeness assessment and quality control are estimated at 15%–40%, representing a substantial economic and nutritional burden, particularly in developing countries [1]. The ripening continuum of a banana—progressing from a starchy, fully green unripe state through mid-ripeness, peak edibility, overripeness, and eventual decomposition—is governed by complex, ethylene-mediated biochemical processes involving starch-to-sugar conversion, chlorophyll degradation with concurrent carotenoid exposure, cell wall softening, and the development of volatile aromatic compounds [2]. Each of these physiological transitions produces distinct and perceptible changes in peel coloration, surface texture, and structural integrity that constitute the primary diagnostic cues exploited by ripeness classification systems.

Conventional ripeness grading in industrial and farm settings relies predominantly on trained human inspectors using standardized color reference charts or tactile evaluation, processes that are inherently subjective, temporally inconsistent, and difficult to scale to large-throughput operations. These limitations motivate the development of automated, objective, and reliable computer vision solutions for ripeness assessment. Convolutional Neural Networks (CNNs) have established the dominant paradigm in agricultural image recognition since their breakthrough performance on the ImageNet large-scale visual recognition challenge [3]. However, CNNs are architecturally predisposed toward local feature extraction through hierarchically stacked convolutional receptive fields, which may constrain their capacity to capture the global contextual cues—such as whole-surface color gradients and spatially distributed browning patterns—that are diagnostically critical for fine-grained ripeness stage differentiation.

Vision Transformers (ViTs), introduced by Dosovitskiy et al. [5], represent a paradigm shift in visual recognition by replacing spatial convolutions with multi-head self-attention mechanisms that inherently model long-range dependencies across non-overlapping image patches. Subsequent architectural innovations, including the hierarchical Swin Transformer with shifted-window attention [6] and the convolution-integrated CvT [7], have addressed the computational scalability and inductive bias limitations of the original ViT, yielding models with improved data efficiency and domain transferability. Despite the proliferating adoption of transformer-based architectures in general computer vision benchmarks, their systematic evaluation against CNNs under rigorous experimental conditions for fine-grained agricultural classification—particularly banana ripeness staging with an inherent ordinal class structure—remains comparatively sparse in the literature.

A further limitation of the existing body of work is the prevalent use of single train–test splits for model evaluation, which may yield optimistic and non-reproducible accuracy estimates.

The ordinal nature of ripeness classes—wherein misclassifying an Unripe banana as Overripe constitutes a qualitatively more severe error than adjacent-class confusion—is seldom quantified through appropriate ordinal error metrics. Additionally, the opacity of deep learning model decisions represents a substantive barrier to trust and adoption in high-stakes agricultural quality control contexts, strongly motivating the integration of Explainable AI (XAI) methods to provide interpretable, visually grounded decision rationales.

This study addresses these gaps through a comprehensive comparative analysis of five deep learning architectures for five-class banana ripeness stage classification, employing a stratified five-fold cross-validation protocol for robust and statistically reliable model selection. Ordinal MAE is incorporated as a complementary evaluation metric alongside standard classification measures. EigenCAM and Grad-CAM are integrated to provide class-discriminative visual explanations for the best-performing model, and the complete classification pipeline is deployed as a production-quality web application to validate end-to-end system viability.

1.1 Contributions

The principal contributions of this work are enumerated as follows:

- A systematic comparative study of CNN (ResNet-50, MobileNetV3-Large) and Vision Transformer (ViT-Base, CvT, Swin Transformer-Tiny) architectures for five-class banana ripeness classification under identical experimental conditions, including identical transfer learning initialization strategy, data augmentation pipeline, optimizer configuration, and hardware environment.
- Application of stratified five-fold cross-validation exclusively to the training partition, yielding statistically robust performance estimates with fold-wise standard deviation analysis to characterize model stability across data subset compositions.
- Integration of Mean Absolute Error (MAE) as an ordinal evaluation metric to characterize the magnitude and severity of misclassifications relative to the ordered ripeness stage structure, providing diagnostic information beyond categorical accuracy.
- Integration of EigenCAM and Grad-CAM post-hoc explainability methods to generate class-discriminative saliency maps for the best-performing model, enabling visual verification that model decisions are grounded in biologically meaningful peel features.
- Design and deployment of a complete classification pipeline as a multi-stage, production-ready Flask web application featuring banana detection validation, real-time background removal, five-fold ensemble inference, and integrated XAI visualization.

2. Literature Review

2.1 Traditional and Conventional Methods for Fruit Grading

Early automated approaches to banana ripeness assessment leveraged handcrafted feature extraction coupled with classical machine learning classifiers. Saad et al. [8] demonstrated that RGB histogram features extracted from banana images, when processed through a multilayer perceptron, achieved approximately 89.3% classification accuracy across three ripeness levels, establishing a foundational baseline for image-based grading. Sabilla et al. [9] employed Principal Component Analysis (PCA) for dimensionality reduction followed by k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), and Decision Tree classifiers, achieving up to 99.1% accuracy for banana type classification; SVM offered the best balance of accuracy and computational efficiency. Sunandar et al. [10] applied the Gray Level Co-occurrence Matrix (GLCM) for texture feature extraction combined with the k-NN algorithm, reaching 90% classification accuracy on a binary ripe/unripe discrimination task. While these approaches are computationally lightweight and offer inherent interpretability by virtue of their handcrafted feature spaces, their generalization capacity is fundamentally constrained by the expressiveness of manually designed features and their sensitivity to imaging variability, particularly changes in illumination, background, and banana cultivar.

2.2 CNN-Based Approaches for Banana Ripeness Classification

The advent of deep convolutional neural networks enabled end-to-end feature learning directly from raw pixel data, substantially advancing agricultural visual recognition tasks. Saragih and Emanuel [3] applied MobileNetV2 with transfer learning from ImageNet to a four-class banana ripeness dataset, achieving 96.18% classification accuracy and demonstrating the efficacy of fine-tuning pre-trained CNN representations for domain-specific agricultural classification. Phoophuangpairoj et al. [11] investigated CNN architectures for classifying whole banana bunches across five ripeness stages, reporting 91.21% accuracy under a 70/30 training–testing split; intra-bunch color variation was identified as a primary source of classification difficulty. Mohamedon et al. [12] developed a mobile application using a TensorFlow Lite CNN that achieved 98.25% accuracy across three ripeness classes, with misclassifications concentrated at the ripe–overripe boundary attributable to spectral similarity.

Object detection frameworks have extended CNN-based banana ripeness systems to localization tasks. Baldovino et al. [13] employed the YOLOv8-S architecture for real-time detection and classification, achieving 89.5% accuracy on static images and 80% accuracy in live webcam testing, illustrating the performance degradation characteristic of real-time deployment. Cahya et al. [14] applied YOLOv5 to multi-object Cavendish banana detection across five

ripeness stages, reporting 98.8% mean average precision. Mutrofin et al. [15] conducted a systematic hyperparameter investigation, finding that simple three-layer CNNs with RMSprop optimization can achieve 99% accuracy on certain datasets, outperforming deeper architectures such as DenseNet201 and VGG16. Thanyacharoen et al. [16] benchmarked five pre-trained CNNs using five-fold cross-validation on a four-class balanced dataset, with VGG16 achieving 93.7% accuracy. While these results confirm the strong representational capacity of CNNs for banana ripeness tasks, they also reveal persistent limitations: limited evaluation rigor, inadequate ordinal analysis, and the absence of model interpretability mechanisms.

2.3 Vision Transformer Approaches in Agricultural Classification

Vision Transformers have demonstrated competitive performance in agricultural image classification tasks, particularly when leveraging large-scale pre-trained representations. Tiwari [17] applied ViT-Base/16 to banana leaf disease classification across seven categories, achieving 96.88% validation accuracy and outperforming EfficientNet-B3, MobileNetV2, and ResNet variants; this study attributed the performance advantage to ViT's capacity for long-range dependency modeling. Xiao et al. [18] compared ViT, Swin Transformer, and MLP for fruit ripeness detection in apple and pear datasets, with the Swin Transformer achieving the highest average precision of 87.43% and the fastest inference time of 0.13 seconds per image, suggesting its particular suitability for small agricultural datasets through hierarchical feature extraction. Fnu and Bansal [29] provided a comprehensive review of ViT variants and their downstream task performance, underscoring the importance of architectural inductive biases for domain-specific adaptation. These results collectively motivate the rigorous evaluation of transformer architectures for banana ripeness staging, where both local chromatic texture discrimination and global surface pattern context are diagnostically relevant.

2.4 Explainable AI in Agricultural Deep Learning

The inherent opacity of deep neural network predictions constitutes a substantive barrier to the trust and regulatory acceptance of automated agricultural systems. Class Activation Mapping (CAM) and its derivatives have emerged as the principal post-hoc tools for generating visual explanations of CNN and transformer model predictions. Gradient-weighted Class Activation Mapping (Grad-CAM) [19], introduced by Selvaraju et al., computes the gradient of a target class score with respect to the feature maps of a designated convolutional layer and applies global average pooling of these gradients to obtain importance weights; the resulting activation maps highlight spatially discriminative image regions. EigenCAM [20], a gradient-free alternative introduced by Muhammad and Yeasin, applies Principal Component Analysis to the intermediate feature maps of a target layer and uses the first principal component—representing the direction of maximum activation variance—as the saliency basis; this approach avoids gradient saturation artifacts that can degrade Grad-CAM quality at high prediction confidence. In the agricultural domain, XAI methods have been applied to leaf disease detection, fruit quality assessment, and pest identification, consistently demonstrating that model attention regions align

with agronomically meaningful features in well-trained models [21]. However, the comparative analysis of EigenCAM versus Grad-CAM for transformer-based banana ripeness classifiers, and their integration within deployed web applications, remains largely unexplored.

2.5 Research Gaps Addressed by This Study

A synthesis of the existing literature reveals five substantive research gaps that collectively motivate the present work. First, the systematic benchmarking of modern hierarchical Vision Transformer architectures against CNNs under rigorously controlled and identical experimental conditions for banana ripeness staging is absent. Second, stratified k-fold cross-validation as a model selection mechanism—essential for reliable performance estimation on moderately sized datasets—is uncommon in this domain. Third, the ordinal structure of ripeness classes is rarely quantified through MAE, which provides more diagnostic information than categorical accuracy for grading applications. Fourth, the comparative evaluation of EigenCAM and Grad-CAM for transformer-based agricultural classifiers, and their integration within deployed systems, is underexplored. Fifth, few published works describe the complete engineering pathway from model training to production deployment within a functional, end-user-accessible application. This study addresses all five gaps within a single unified experimental framework.

3. Methodology

The proposed experimental framework integrates dataset curation, preprocessing, multi-architecture model training under cross-validation, XAI integration, and web application deployment within a unified pipeline. The complete system architecture is illustrated in Fig. 1.

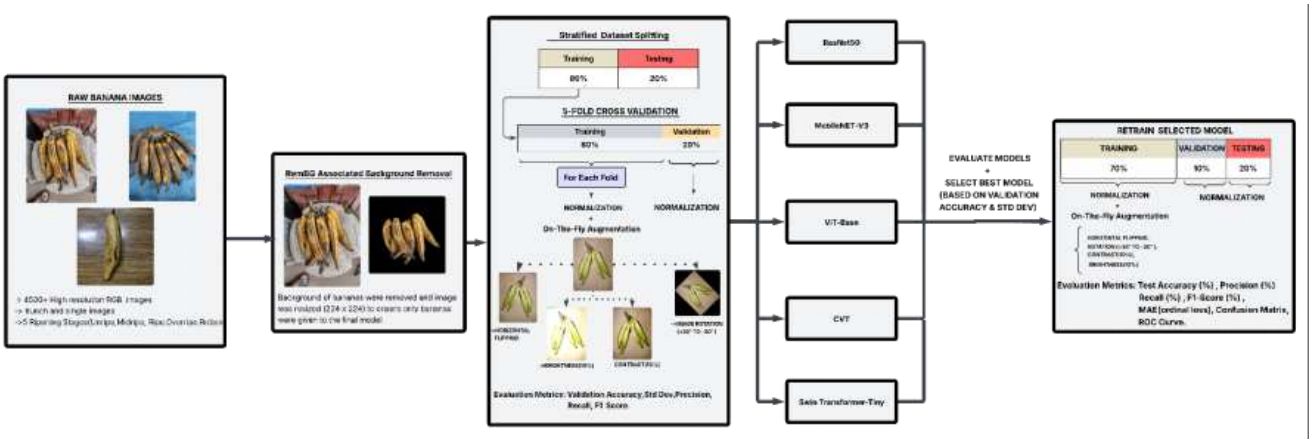


Fig. 1. Complete system methodology pipeline: from raw banana image collection and background removal through stratified five-fold cross-validation across five architectures, to final model retraining, evaluation, and deployment.

3.1 Dataset Description

The study utilized a curated dataset of 4,500 RGB images encompassing five banana ripeness stages: Unripe, Mid-Ripe, Ripe, Overripe, and Rotten. This five-class taxonomy covers the complete post-harvest ripening trajectory relevant to commercial grading, retail quality control, and food waste reduction. Each class contains exactly 900 images, ensuring a balanced class distribution that eliminates class imbalance as a confounding variable in comparative performance evaluation. Images were collected across diverse imaging conditions—including varied ambient illumination intensities, background environments, camera viewing angles, and shooting distances—to maximize the ecological validity of the dataset and enhance model generalization. Both individual bananas and bunch formations were included to reflect the diversity of real-world inspection scenarios. The dataset was partitioned into a training set (3,600 images, 80%) and a hold-out test set (900 images, 20%); the test set was withheld from all model selection procedures to ensure an unbiased final performance estimate.

3.2 Data Preprocessing and Augmentation

All images were subjected to a three-stage preprocessing pipeline prior to model ingestion. First, background removal was performed using the RemBG library, which applies a pre-trained salient object detection network to generate alpha-masked foreground extractions, effectively isolating the banana specimen from its background. This step is particularly important because the chromatic properties of certain background environments can be confounded with ripeness-indicative peel colors under naive pixel-level analysis. As illustrated in the system pipeline (Fig. 1), background removal ensures that downstream feature extraction is focused exclusively on the banana subject. Second, all foreground-masked images were resized to 224×224 pixels to conform to the input dimensionality required by the ImageNet pre-trained backbone weights used for transfer learning initialization across all five architectures. Third, per-channel normalization was applied using the channel-wise mean and standard deviation computed from the training partition, centering the input distribution to stabilize gradient flow and accelerate convergence during fine-tuning.

On-the-fly data augmentation was applied exclusively to training subset samples during model fitting to improve generalization without inflating the training set size through offline pre-augmentation. The augmentation pipeline comprised: random horizontal flipping with probability 0.5, random rotation within ±10 degrees, and mild color jitter with brightness and contrast perturbations bounded to ±10% of original values. The deliberately conservative magnitude of color jitter was motivated by the need to preserve the chromatic integrity of ripeness-stage peel coloration, which constitutes the primary discriminative signal; excessive

color perturbations risk corrupting this signal and introducing misleading training gradients. No augmentation was applied to validation or test partitions.

3.3 Model Architectures

Five deep learning architectures spanning a representative range of inductive biases and parameter scales were selected for comparative evaluation. All models were initialized with ImageNet-1K pre-trained weights to leverage large-scale visual representations, and fine-tuned under a partial parameter update strategy in which the bottom 75% of parameters were frozen and only the top 25% of parameters—including the task-specific classification head—were updated during training. This strategy balances the preservation of generalizable pre-trained representations with the need for domain-specific adaptation while mitigating overfitting on the moderately sized training set.

ResNet-50 [22] is a 50-layer residual network comprising four stages of bottleneck residual blocks with progressively increasing channel depth (64→512), followed by global average pooling and a fully connected classification head. Skip connections introduced in each residual block enable effective gradient propagation through depth, addressing the vanishing gradient problem that limits very deep plain CNNs. For the five-class banana task, the final linear layer was replaced with a 2,048→5 dimensional projection. MobileNetV3-Large [23] is a compact CNN designed for mobile deployment through depthwise separable convolutions, hard-swish activations, and channel-wise Squeeze-and-Excitation (SE) attention modules. Its inverted residual architecture achieves competitive accuracy with substantially reduced parameter count, making it attractive for resource-constrained inference environments. The classification head was adapted to produce five class logits while retaining the preceding activation and dropout layers.

ViT-Base/16 [5] partitions the input image into non-overlapping 16×16 pixel patches, linearly projects each patch to a 768-dimensional token embedding, and processes the resulting token sequence through 12 standard Transformer encoder blocks, each comprising multi-head self-attention and feed-forward sub-layers with Layer Normalization. A learnable class token is prepended to the patch sequence, and its final representation serves as the image embedding for classification. The Convolutional Vision Transformer (CvT) [7] augments this architecture by replacing linear patch projections with convolutional token embeddings and incorporating depthwise separable convolutions in the query, key, and value projections of the attention mechanism, introducing local spatial inductive bias while preserving global attention capacity.

The Swin Transformer [6] introduces a hierarchical, shifted-window self-attention mechanism that processes feature maps through four stages of increasing channel depth and decreasing spatial resolution, analogously to CNN feature pyramids. Within each stage, self-attention is computed within non-overlapping local windows, with alternating layers employing cyclically shifted window partitioning to establish cross-window connectivity. This design reduces the computational complexity of self-attention from quadratic to linear in image

size while enabling multi-scale feature representations that capture both fine-grained local texture and broader spatial context. The Swin-Tiny variant (28M parameters) was employed in this study, with the classification head replaced by a five-class linear projection.

3.4 Training Strategy and Cross-Validation Protocol

All five models were trained under identical hyperparameter configurations to ensure fair comparative evaluation. The AdamW optimizer [24] was employed with an initial learning rate of 3×10^{-4} and default weight decay regularization ($\lambda = 0.01$). Categorical cross-entropy was used as the training loss function. Training proceeded for a maximum of 50 epochs per fold, with early stopping triggered upon five consecutive epochs without improvement in validation loss. A batch size of 16 was used across all experiments. All training was conducted on a GPU-accelerated computing environment.

A stratified five-fold cross-validation scheme was applied exclusively to the 3,600-image training partition. Stratification ensures that each fold maintains the original balanced class distribution (180 samples per class per fold), preventing distributional shifts between training and validation subsets. For each fold, 2,880 images (80% of the training partition) served as the fitting set and 720 images (20%) as the validation set. Model selection was based on the fold-averaged validation accuracy and its standard deviation across folds, with lower standard deviation signaling greater stability. Following cross-validation model selection, the winning architecture was retrained on the complete 3,600-image training partition with an internal 80:20 split for learning curve monitoring, then evaluated on the held-out 900-image test set for the final unbiased performance estimate.

3.5 XAI Integration: Grad-CAM and EigenCAM

To provide post-hoc visual explanations for the predictions of the selected model, two complementary class activation mapping techniques were implemented using the PyTorch Grad-CAM library. Grad-CAM [19] computes the gradient of the target class score with respect to the feature maps of a designated target layer, applies global average pooling of these gradients to obtain layer-specific channel importance weights, and computes a weighted combination of the feature maps followed by ReLU activation to isolate positive contributions. For the Swin Transformer, Grad-CAM was applied to the output feature maps of the final hierarchical stage, corresponding to the deepest semantic representation preceding the classification head. EigenCAM [20] applies Principal Component Analysis to the feature maps of the same target layer and uses the first principal component as the saliency basis, providing a gradient-free explanation that is robust to gradient saturation artifacts. Both methods generate heatmaps overlaid on the original input image using a jet colormap, with red indicating high activation relevance and blue indicating low relevance.

3.6 Evaluation Metrics

Model performance was assessed using a comprehensive suite of metrics appropriate for multi-class classification with ordinal class structure. Categorical accuracy, macro-averaged precision, recall, and F1-score were computed from the confusion matrix on the held-out test set; macro-averaging assigns equal weight to all five classes, preventing dominant-class performance from masking per-class deficiencies. The area under the Receiver Operating Characteristic curve (AUC-ROC) was computed per class in a one-versus-rest formulation to assess discriminative capacity across varying classification thresholds. Mean Absolute Error (MAE) was computed as an ordinal metric, quantifying the average distance in ripeness stage steps between predicted and true labels, with larger values indicating more severe misclassifications relative to the ordinal class ordering. Cross-validation standard deviation was reported as a model stability indicator.

4. Results and Discussion

4.1 Cross-Validation Performance Comparison

Table 1 presents the fold-wise and aggregate validation accuracies, along with macro precision, recall, F1-score, and standard deviation, for all five architectures across the stratified five-fold cross-validation protocol. The Swin Transformer consistently achieves the highest per-fold validation accuracies, ranging from 86.53% (Fold 5) to 88.75% (Fold 4), yielding a mean accuracy of 87.39% with the lowest observed standard deviation of 0.868%. These results indicate that the Swin Transformer not only achieves superior average performance but also generalizes more reliably across different training–validation data partitions, a critical property for deployment in variable real-world imaging conditions.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg Acc (%)	Std Dev	Avg P	Avg R	Avg F1
ViT-Base	86.96	85.69	85.56	86.94	84.17	85.86	1.157	0.8616	0.8586	0.8584
CvT	85.69	85.56	84.58	87.36	84.86	85.61	1.083	0.8601	0.8561	0.8551
Swin Transformer	86.81	87.22	87.63	**88.75**	86.53	**87.39**	**0.868**	**0.8765**	**0.8740**	**0.8734**
MobileNetV3	82.92	83.75	82.50	86.25	84.17	83.92	1.461	0.8412	0.8391	0.8380
ResNet-50	82.64	87.92	84.58	86.39	84.17	85.14	2.049	0.8544	0.8513	0.8502

Table 1. Extended cross-validation performance of all architectures (stratified 5-fold).

The bar chart in Fig. 2 visually confirms the performance hierarchy, with the Swin Transformer bar clearly elevated above all competitors and the shortest error bar, indicating the lowest cross-fold variability. Both ViT-based variants (ViT-Base: 85.86%, CvT: 85.61%)

outperform both CNN baselines in average accuracy, suggesting that attention-based global context modeling provides a genuine advantage for banana ripeness classification over purely local convolutional feature extraction. The box plot comparison in Fig. 3 provides additional insight into distributional variability: the Swin Transformer's interquartile range is notably compact and its median is highest, while ResNet-50 exhibits the widest spread and a visible outlier corresponding to Fold 2 (87.92%), which likely reflects the particularly favorable class distribution alignment in that specific fold. MobileNetV3, despite its computational efficiency advantages, achieves the lowest mean accuracy of 83.92%, illustrating the fundamental accuracy–efficiency trade-off in compact CNN design.

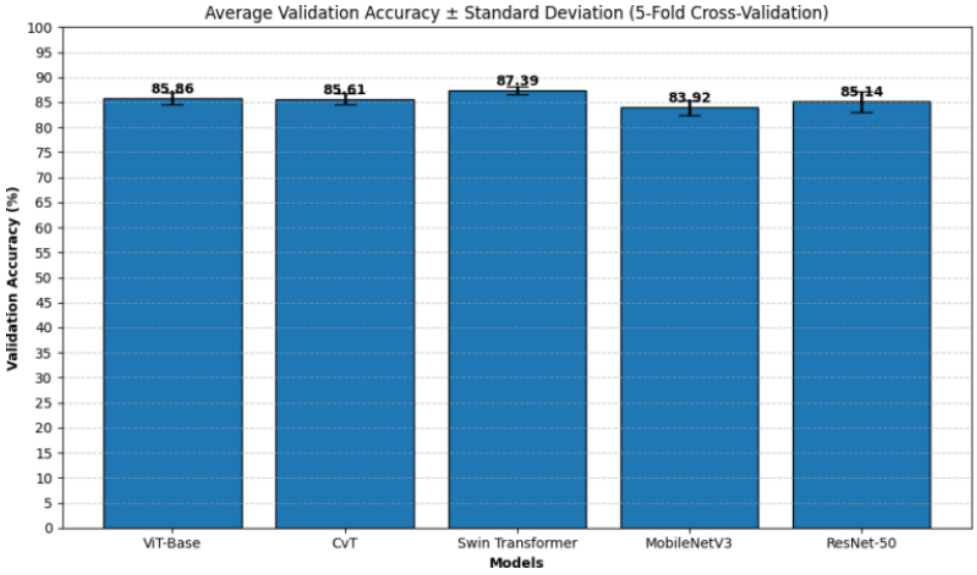


Fig. 2. Average validation accuracy (%) ± standard deviation across all five architectures under stratified five-fold cross-validation. The Swin Transformer achieves the highest mean accuracy (87.39%) and the smallest error bar, indicating superior and most consistent generalization.

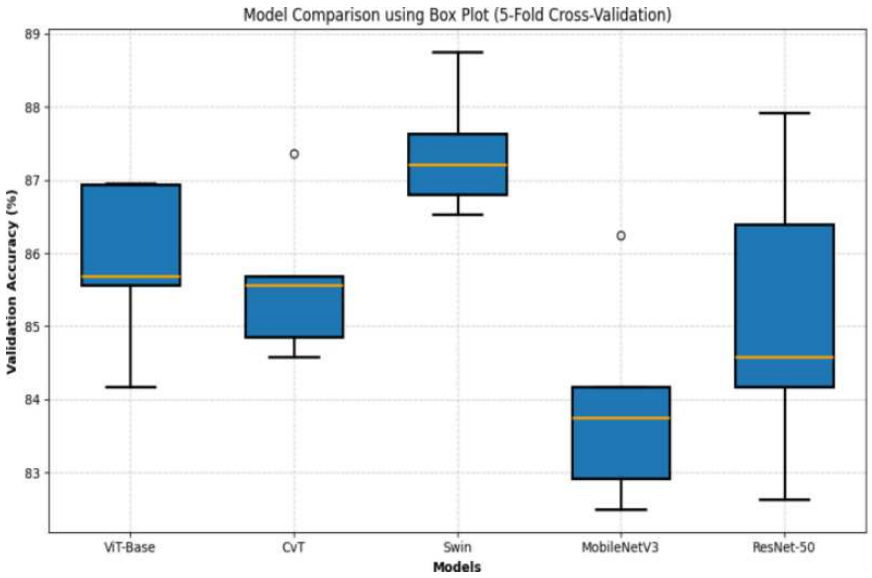


Fig. 3. Box plot of fold-wise validation accuracy distributions for all five models (5-fold cross-validation). The Swin Transformer displays the highest median and most compact interquartile range, confirming both performance superiority and stability across data partitions.

4.2 Training Dynamics of the Selected Model

The training and validation accuracy and loss curves for the final Swin Transformer model retrained on the complete training partition are shown in Fig. 4 and Fig. 5, respectively. Both curves exhibit the expected pattern of rapid initial improvement followed by gradual convergence, consistent with effective transfer learning initialization. The training accuracy curve (Fig. 4) rises steeply in the first ten epochs—from approximately 55% to 80%—as the classification head and unfrozen upper layers rapidly adapt to the banana ripeness feature space, then continues to improve steadily, plateauing near 86% by epoch 50. The validation accuracy trajectory closely tracks the training curve throughout training, reaching approximately 84%, with no significant divergence between the two curves. This parallel behavior indicates that the model is well-regularized and does not exhibit significant overfitting, attributable to the combination of early stopping, conservative augmentation, and the partial parameter freezing strategy.

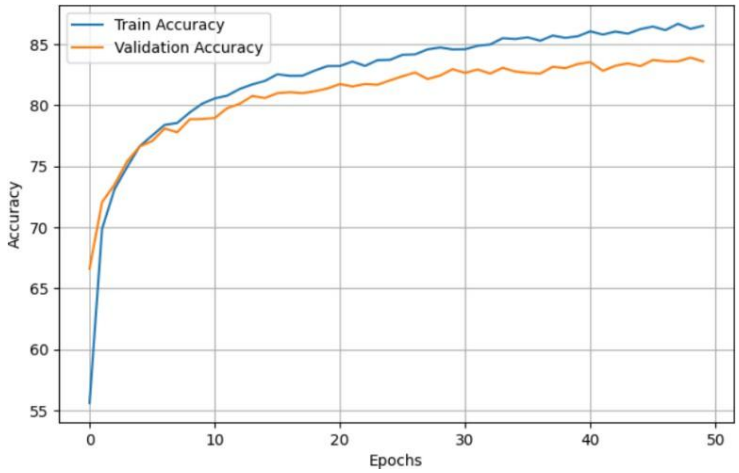


Fig. 4. Training and validation accuracy curves of the Swin Transformer (final model, 50 epochs). The close tracking of training and validation accuracy throughout training indicates well-regulated generalization without significant overfitting.

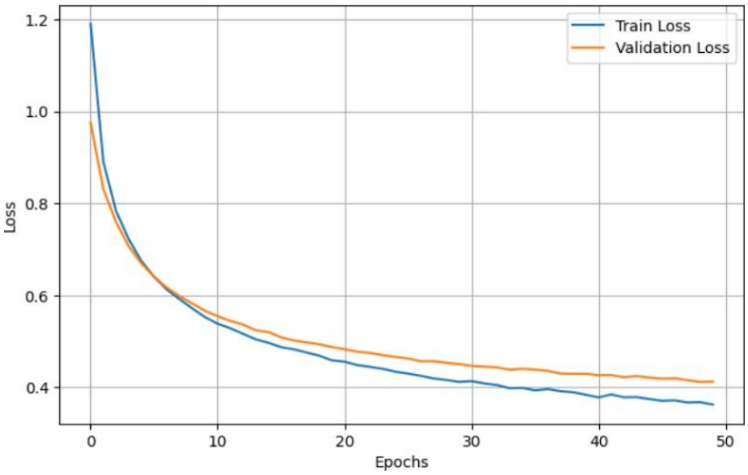


Fig. 5. Training and validation loss curves of the Swin Transformer (final model, 50 epochs). The smooth, monotonically decreasing loss trajectories with minimal validation–training gap confirm effective convergence and absence of overfitting.

The loss curves (Fig. 5) demonstrate smooth, monotonically decreasing trajectories for both training and validation losses, converging from approximately 1.2 and 0.98 at epoch 0 to approximately 0.36 and 0.42 at epoch 50, respectively. The consistent gap of approximately 0.06 loss units between training and validation at convergence is modest and stable, further corroborating the absence of harmful overfitting. The absence of sudden divergence or oscillation in the validation loss confirms that the early stopping patience of five epochs was not prematurely triggered and that the final model checkpoint represents a well-converged solution.

4.3 Final Test Set Evaluation and Class-Wise Analysis

Table 2 presents the class-wise and aggregate classification report for the Swin Transformer evaluated on the 900-image hold-out test set. The model achieves an overall test accuracy of 86.78% with macro precision of 0.88, recall of 0.87, and F1-score of 0.87. The confusion matrix (Fig. 6) and class-wise performance bar chart (Fig. 7) provide complementary granular views of these results.

Class	Precision	Recall	F1-Score	Accuracy (%)	Support
Unripe	0.95	0.79	0.86	78.89	180
Mid-Ripe	0.74	0.95	0.83	95.00	180
Ripe	0.88	0.83	0.86	83.33	180
Overripe	0.91	0.79	0.85	79.44	180
Rotten	0.90	0.97	0.94	97.22	180
Macro Average	**0.88**	**0.87**	**0.87**	**86.78**	**900**

Table 2. Class-wise and aggregate test-set performance of the Swin Transformer on 900 test images.

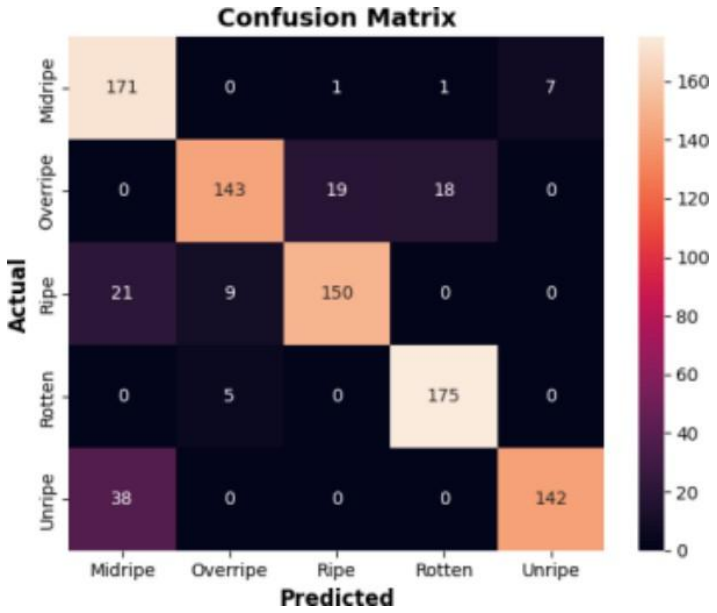


Fig. 6. Normalized confusion matrix of the Swin Transformer on the 900-image test set. Diagonal dominance confirms strong per-class discriminative capacity; off-diagonal errors are concentrated in adjacent class pairs, consistent with the ordinal structure of ripeness stages.

Analysis of the confusion matrix (Fig. 6) reveals that all five classes exhibit strong diagonal dominance, confirming that the model correctly classifies the majority of instances in each category. The Rotten class achieves the highest classification accuracy (97.22%), with 175 out of 180 test samples correctly identified. The confusion matrix confirms that the five misclassified

Rotten samples were assigned to the Overripe class—the adjacent stage—rather than to distal classes, consistent with the progressive browning continuum between these stages. Similarly, the Mid-Ripe class achieves 95% accuracy (171/180), with one sample each misclassified as Ripe and Rotten, and seven misclassified as Unripe. The Unripe class (142/180, 78.89%) and Overripe class (143/180, 79.44%) exhibit the highest error rates, with Unripe samples predominantly misclassified as Mid-Ripe (38 samples) and Overripe samples distributed between Ripe (19) and Rotten (18).

These patterns are agronomically interpretable: the transition from Unripe to Mid-Ripe involves a gradual green-to-yellow peel color shift with individual specimen variability in rate, creating chromatic ambiguity at the class boundary. The Overripe–Rotten boundary similarly lacks a sharp visual discontinuity, as peel browning progresses continuously with tissue decomposition. The class-wise performance comparison chart in Fig. 7 visually confirms that the Rotten class achieves the most balanced and highest performance profile across all four metrics, while Mid-Ripe and Overripe exhibit the most notable metric asymmetries—high recall but low precision for Mid-Ripe, and high precision but lower recall for Overripe—reflecting the directional nature of confusion between adjacent stages.

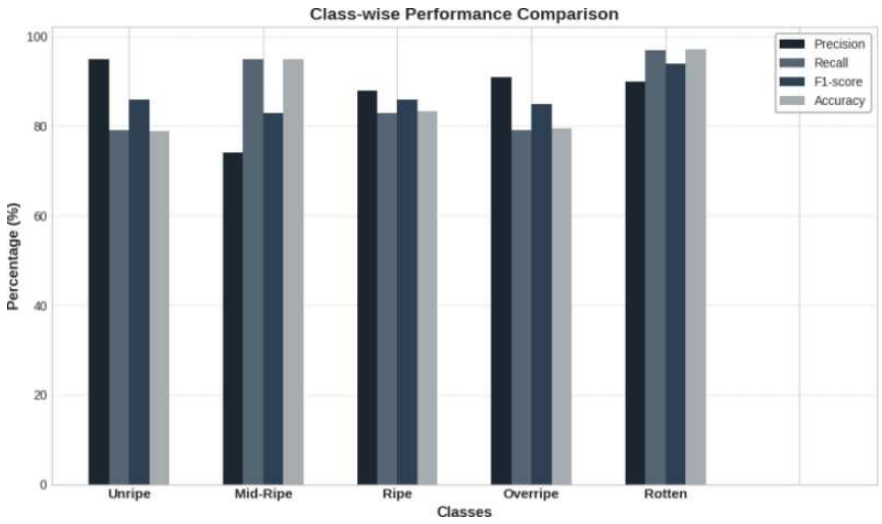


Fig. 7. Class-wise performance comparison for the Swin Transformer (precision, recall, F1-score, and accuracy). The Rotten class achieves the most balanced and highest metrics; asymmetries in Mid-Ripe and Overripe reflect the directional nature of confusion at adjacent ripeness stage boundaries.

4.4 Ordinal Error Analysis

The computed ordinal MAE of 0.3344 over the 900-image test set confirms that the Swin Transformer's misclassifications are predominantly concentrated between immediately adjacent ripeness stages. Since each unit of MAE corresponds to a one-stage ordinal distance (e.g., predicting Overripe when the true class is Rotten corresponds to MAE = 1), an MAE of 0.3344

indicates that the weighted average prediction error represents less than one-third of a stage displacement. This result quantitatively validates the observation from the confusion matrix that off-diagonal errors are confined to neighboring class cells, confirming that the model has acquired an implicit understanding of the ordinal structure of the ripeness progression without explicit ordinal supervision. From a practical agricultural quality assurance standpoint, an MAE below 0.5 stages indicates that virtually all errors are agronomically acceptable single-stage misassignments, which are significantly less consequential than multi-stage errors that could, for example, lead to the retail release of Rotten produce or the premature rejection of perfectly Ripe stock.

4.5 Discriminative Capacity and ROC Analysis

The Receiver Operating Characteristic (ROC) curves for all five ripeness classes, computed in a one-versus-rest formulation on the test set, are shown in Fig. 8. All five classes achieve AUC values exceeding 0.98, with the Rotten class attaining a near-perfect AUC of 1.00. The exceptionally high discriminative capacity reflected in these AUC values demonstrates that the Swin Transformer's output probability estimates are well-calibrated across a wide range of classification thresholds, not just at the default argmax threshold. The Rotten class's perfect separation confirms that no test sample from another class is assigned a higher Rotten probability than any true Rotten sample, reflecting the visual distinctiveness of fully decomposed banana specimens. The Unripe class achieves AUC = 0.99, while Mid-Ripe, Overripe, and Ripe all achieve AUC = 0.98, indicating uniformly excellent discriminability across all classes despite the accuracy asymmetries observed in the confusion matrix analysis.

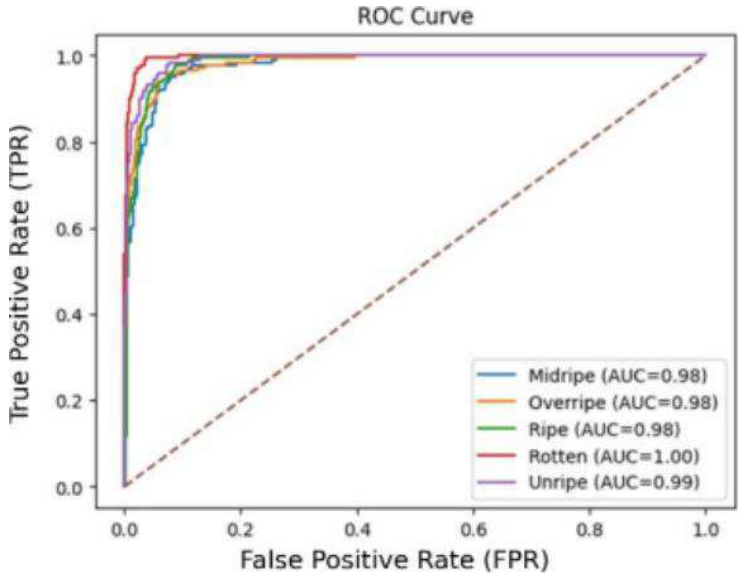


Fig. 8. Receiver Operating Characteristic (ROC) curves and area under the curve (AUC) values for all five ripeness classes (one-versus-rest evaluation on 900 test images). All classes achieve AUC ≥ 0.98, confirming excellent discriminative capacity across classification thresholds.

The consistently high AUC values observed across all classes indicate that the strong macro-averaged test accuracy of 86.78% is not an artifact of performance concentration in easily separable classes; rather, the model maintains robust discriminability for all five classes, including the transitional Mid-Ripe and boundary-adjacent Overripe classes. This finding strengthens the case for the Swin Transformer as a reliable solution for operational banana quality grading.

4.6 Comparative Summary and Architectural Analysis

Table 3 provides a consolidated architecture comparison based on cross-validation performance.

Model	Avg Acc (%)	Std Dev (%)	Architectural Interpretation
ViT-Base	85.86	1.157	Good stability; global attention without multi-scale hierarchy
CvT	85.61	1.083	Good stability; convolutional inductive bias mitigates patch boundary discontinuities
Swin Transformer	**87.39**	**0.868**	Best: hierarchical shifted-window attention captures multi-scale ripeness cues with low variance
MobileNetV3	83.92	1.461	Lowest accuracy; SE modules partially compensate for limited depth
ResNet-50	85.14	2.049	Highest fold variance; purely local convolutions sensitive to color distribution shifts

Table 3. Summary comparison of all architectures: cross-validation accuracy–stability trade-off.

The performance ranking observed across the five architectures is consistent with theoretical expectations derived from their respective inductive biases. The Swin Transformer's advantage over both ViT-Base and CvT suggests that hierarchical multi-scale representation—analogous to CNN feature pyramids but implemented through self-attention mechanisms—is particularly beneficial for banana ripeness classification, where discriminative cues operate at multiple spatial scales: fine-grained surface texture (spotting, wrinkling) at small scales and global peel

color distribution at large scales. ViT-Base's competitive but lower accuracy relative to Swin reflects the inherent limitation of flat, non-hierarchical attention in simultaneously representing features at multiple granularities without explicit scale-space structure. CvT's convolutional embeddings provide a useful local inductive bias but do not replicate the hierarchical multi-scale design of Swin. ResNet-50's high fold-to-fold variance is consistent with the known sensitivity of deep CNNs to the specific color and texture distributions encountered in training batches, particularly for color-critical classification tasks.

5. Explainable AI Analysis

5.1 Visual Saliency Mapping with Grad-CAM and EigenCAM

To elucidate the spatial feature attribution patterns underlying the Swin Transformer's predictions and verify their agronomic validity, Grad-CAM and EigenCAM saliency maps were computed for representative test images from each ripeness class. Fig. 9 presents the complete XAI visualization panel for a representative Rotten class sample, illustrating the raw input image, the background-removed preprocessing output, the Grad-CAM (ViTGradCAM) overlay, the EigenCAM overlay, and the Attention Rollout visualization, which traces the cumulative attention propagation through all transformer layers from patch tokens to the class token.

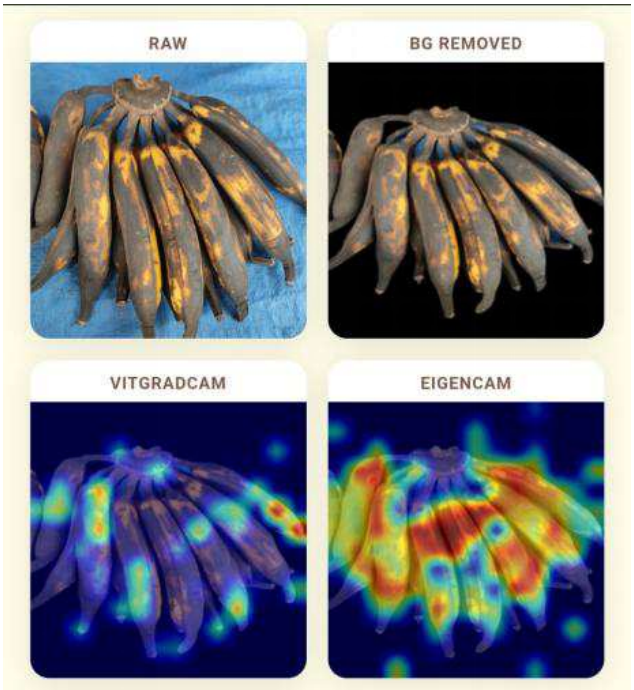


Fig. 9. XAI visualization panel for a representative Rotten-class banana: (a) raw input image; (b) background-removed preprocessing output; (c) Grad-CAM (ViTGradCAM) saliency overlay; (d) EigenCAM saliency overlay. Red regions indicate highest model activation; blue indicates lowest activation.

For the Rotten class example shown in Fig. 9, Grad-CAM generates discrete high-activation spots concentrated on the peel regions exhibiting the most severe dark browning and surface decomposition, as evidenced by the red hotspots distributed across the proximal and distal peel areas of the banana bunch. This spatially localized attribution is consistent with the diagnostic importance of concentrated tissue decomposition indicators for the Rotten class, where globally uniform browning produces multiple co-equal focal points of classification-relevant evidence. EigenCAM, in contrast, generates a broader activation distribution spanning the entire anterior surface of the peel, with maximum activation (red) concentrated on the central dorsal region of the bunch where browning is most extensive. This complementary broad coverage reflects EigenCAM's basis in the first principal component of the full feature map, which captures the global directional trend of activation rather than the localized peak responses identified by gradient weighting.

The Attention Rollout visualization reveals a diffuse but predominantly foreground-focused attention pattern, confirming that the background-removed preprocessing has successfully suppressed spurious background token activations. The banana bunch receives substantially higher cumulative attention than the padding regions in all transformer layers, validating that the model's hierarchical shifted-window attention mechanism effectively focuses on the diagnostically relevant foreground object throughout the depth of the network.

5.2 Class-Wise Saliency Interpretation

Across all five ripeness classes, the saliency maps generated by both Grad-CAM and EigenCAM exhibit consistent alignment with the biologically expected diagnostic features of each stage. For the Unripe class, both methods concentrate activations on the uniformly green proximal peel regions, consistent with the chlorophyll-dominated spectral signature characteristic of pre-climacteric bananas prior to ethylene-mediated pigmentation changes. For the Mid-Ripe class, activation regions shift toward the peel tips and curvature areas—the anatomical sites where yellow-green color transition initiates—reflecting the model's sensitivity to the earliest chromatic indicators of the climacteric transition. For the Ripe class, high-activation regions span the uniformly yellow peel surface with peak activations at the distal tips, consistent with completed chlorophyll degradation and full carotenoid expression.

For the Overripe class, Grad-CAM highlights discrete brown spotting distributed across the peel—characteristic of the superficial tissue damage and enzymatic browning that emerges in post-climacteric bananas—while EigenCAM captures the overall yellow-to-brown color gradient spanning the peel surface. This complementarity is particularly valuable for the Overripe class because the diagnostic signal is distributed rather than locally concentrated, and neither method alone provides a complete characterization. For the Rotten class (as illustrated in Fig. 9), both methods produce high-activation maps covering the majority of the visible peel surface,

reflecting the globally consistent dark coloration and structural degradation that characterize advanced decomposition. The observation that model activations faithfully reflect these class-specific physiological indicators—rather than spurious background or non-diagnostic image artifacts—provides strong evidence that the Swin Transformer has learned biologically meaningful representations aligned with the agronomic understanding of banana ripening.

5.3 Comparative Analysis of EigenCAM and Grad-CAM

A systematic comparison of the two XAI methods reveals complementary strengths that are well-suited to different aspects of banana ripeness interpretation. Grad-CAM consistently produces sharper, more spatially localized activation maps with higher contrast between salient and non-salient regions, making it effective for pinpointing specific discriminative spots such as browning foci in Overripe bananas or the green tips of Unripe specimens. EigenCAM generates smoother, broader activations with more diffuse spatial support, better capturing the holistic peel color distribution that dominates the Ripe and Rotten stage diagnostics. This gradient-free approach is also demonstrably more robust in the high-confidence prediction regime—exemplified by the Rotten class (96.51% confidence)—where gradient saturation can cause Grad-CAM heatmaps to lose contrast or exhibit uninformative uniform distributions. Taken together, the two methods provide mutually reinforcing evidence: Grad-CAM identifies what specific features trigger the prediction, while EigenCAM characterizes the overall activation pattern that the model associates with the predicted class.

6. System Implementation

6.1 Web Application Architecture

The complete classification pipeline was implemented as a production-quality web application using the Flask micro-framework to validate end-to-end system viability and practical deployability. The application design incorporates three principal processing stages that mirror the model training pipeline: multi-stage banana detection validation, background removal preprocessing, and ensemble classification with integrated XAI visualization.

The banana detection validation stage employs a YOLOv8-nano object detector as a primary verification gate, confirming the presence of a banana-like object in the submitted image with a configurable confidence threshold. This gate is supplemented by a complementary dual-metric fallback mechanism operating in HSV color space: a color compatibility score assessing the chromatic similarity of the dominant foreground pixels to the expected ripeness-stage color palette, and a contour-based morphological score evaluating the elongation ratio, solidity, and area coverage of detected foreground contours. This fallback mechanism is specifically designed to handle Overripe and Rotten bananas, for which YOLO detection confidence may be reduced

due to the visual dissimilarity of these specimens from the banana training distribution of general-purpose object detectors. Only images passing both the primary detection gate and the fallback validation are forwarded to the classification pipeline, preventing erroneous ripeness predictions for non-banana inputs.

Following validation, background removal is applied using RemBG, replicating the preprocessing pipeline used during model training to ensure distributional consistency between training and inference data. Classification is then performed through five-fold ensemble inference: each of the five cross-validation fold model checkpoints independently produces a class probability vector, and the five probability vectors are aggregated through majority voting on their respective argmax predictions to produce the final ensemble ripeness classification. This ensemble strategy reduces prediction variance relative to any individual fold model by leveraging the diversity of decision boundaries learned from different training data compositions.

The application interface, shown in Fig. 10, provides a user-friendly image upload panel with Analyze and Clear controls, an immediately visible ripeness prediction with confidence score, and a multi-fold ensemble vote indicator. The main results panel (Fig. 11) displays the full class probability distribution as an interactive bar chart alongside a detailed classification explanation card identifying the predicted class, confidence level, and natural-language description of the visual features detected by the model. All model weights for the five cross-validation folds are stored on and automatically downloaded from the Hugging Face Hub at application initialization, enabling stateless containerized deployment without manual weight file management.



Fig. 10. Main interface of the deployed Flask-based banana ripeness web application. The user uploads a banana image, initiates analysis, and receives a ripeness classification (here: 'Rotten', 96.51% confidence) with a five-fold ensemble prediction indicator.

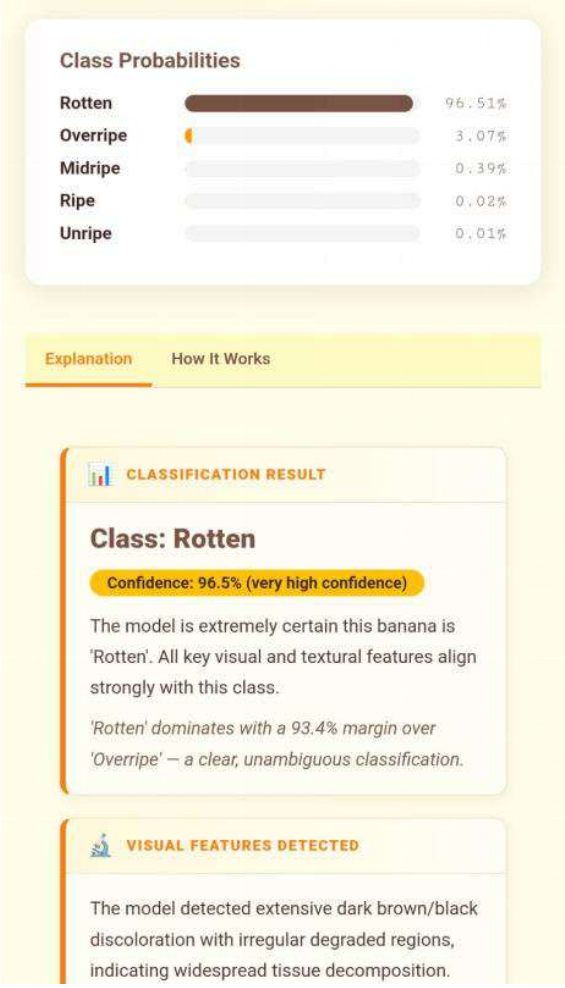


Fig. 11. Classification results panel of the web application displaying: (a) class probability distribution for the Rotten prediction (96.51% confidence); (b) natural-language classification explanation card identifying detected visual features and prediction margin over the nearest class.

6.2 Multi-Class Inference Examples

Fig. 12 presents the web application interface across three representative inference examples spanning the ripeness spectrum, demonstrating the system's operational capability on diverse input images. The Rotten class example (Fig. 12a) yields 96.51% ensemble confidence, reflecting the high visual distinctiveness of advanced decomposition. The Ripe class example (Fig. 12b) yields 76.35% confidence, appropriately lower given the higher visual similarity of Ripe bananas to adjacent Mid-Ripe and Overripe stages; this moderate confidence correctly reflects the model's uncertainty at the more ambiguous transitional stage boundary. The Unripe class example (Fig. 12c) yields 81.96% confidence, reflecting the relatively distinctive appearance of fully green pre-climacteric bananas with only marginal confusion risk from Mid-Ripe samples at the earliest color transition stage.

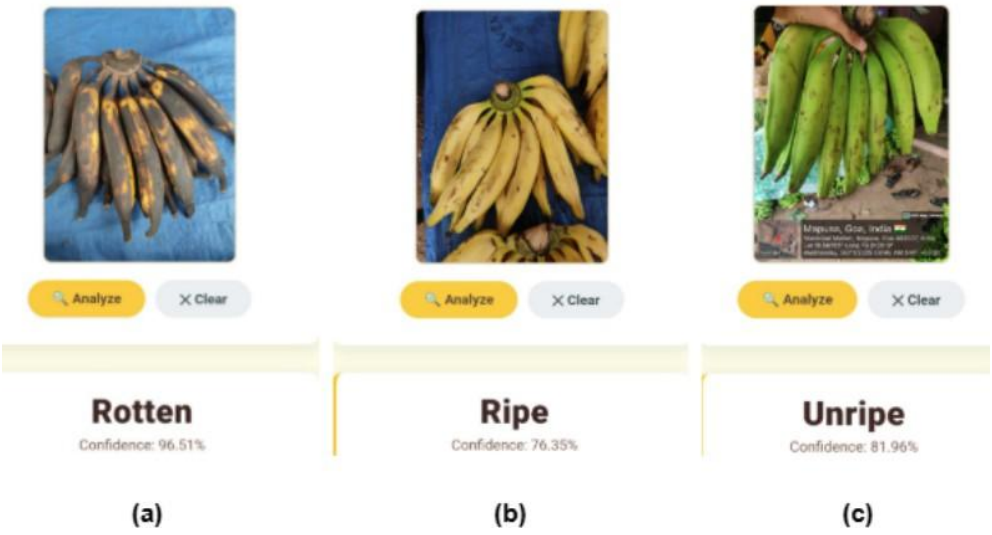


Fig. 12. Representative web application inference results across three ripeness classes: (a) Rotten — 96.51% confidence; (b) Ripe — 76.35% confidence; (c) Unripe — 81.96% confidence. Confidence levels reflect the relative visual distinctiveness of each class relative to its neighbors.

The pattern of confidence scores across these examples is agronomically consistent: classes with the most distinctive and globally homogeneous visual appearance (Rotten, Unripe) yield higher confidence predictions, while classes near the middle of the ordinal spectrum where spectral similarity to neighbors is greatest (Ripe) yield lower but still correct predictions. This calibrated confidence behavior indicates that the ensemble inference strategy produces well-calibrated probability estimates that can support threshold-based decision-making in quality control applications, where the cost of false positive and false negative predictions may differ across class pairs.

6.3 Deployment Infrastructure

The application is deployed on Hugging Face Spaces using a Docker container with CUDA-optional inference support, enabling GPU-accelerated inference when available and graceful fallback to CPU-only operation for cost-effective deployment. The five Swin Transformer fold checkpoints (each approximately 110 MB) are hosted on the Hugging Face Hub as model repository artifacts and downloaded to the container's ephemeral storage at initialization using the hub download API. This design decouples model storage from application code, simplifying version management, enabling checkpoint updates without container rebuilds, and minimizing container image size. The total application initialization time, including model downloads and weight loading, is approximately 45 seconds on a CPU-only inference environment, with per-image inference time—including background removal, ensemble prediction, and dual XAI map generation—averaging approximately 8–12 seconds.

7. Conclusion

This paper presented a comprehensive comparative evaluation of five deep learning architectures—ResNet-50, MobileNetV3-Large, ViT-Base, CvT, and Swin Transformer-Tiny—for automated five-class banana ripeness stage classification. A curated 4,500-image dataset with systematic background removal and conservative augmentation was used as the experimental testbed. Stratified five-fold cross-validation provided robust, low-variance model selection criteria, and ordinal MAE supplemented standard classification metrics to characterize misclassification severity relative to the ordered ripeness taxonomy.

The Swin Transformer emerged as the superior architecture in terms of both mean accuracy (87.39% cross-validation) and stability (standard deviation 0.868%), outperforming all CNN and non-hierarchical transformer competitors. The final test evaluation confirmed a test accuracy of 86.78% with macro F1-score of 0.87 and AUC values exceeding 0.98 across all five ripeness classes. The ordinal MAE of 0.3344 validated that residual prediction errors are confined to adjacent ripeness stage boundaries, confirming taxonomic consistency appropriate for quality grading applications. Grad-CAM and EigenCAM visual explanations produced biologically meaningful, class-consistent saliency maps aligned with the peel coloration, spotting patterns, and browning gradients characteristic of each ripeness stage, providing interpretable evidence that the model's predictions are grounded in agronomically relevant visual features rather than spurious artifacts.

The complete system was successfully deployed as a Flask-based web application with multi-stage banana validation, five-fold ensemble inference, and integrated XAI visualization, demonstrating end-to-end practical viability for precision agriculture. Future research directions include: expanding the dataset to encompass diverse Musa species and cultivars across uncontrolled field and retail imaging environments; investigating larger Swin Transformer configurations (Small, Base) and knowledge distillation techniques to produce compact yet accurate variants for edge-device deployment; extending the classification framework to detection-level multi-banana localization and individual-specimen ripeness estimation within bundle images using detection-transformer architectures; exploring multi-modal sensing fusion incorporating near-infrared, hyperspectral, or gas-sensor modalities to capture post-harvest ripeness indicators not visible in RGB imaging alone; and incorporating temporal ripening trajectory modeling through recurrent or temporal convolutional networks to support predictive shelf-life management and dynamic cold-chain logistics optimization.

References

[1] Food and Agriculture Organization of the United Nations, "FAOSTAT: Crops and Livestock Products," FAO, Rome, 2023. [Online]. Available: https://www.fao.org/faostat

[2] D. O. Ekramirad, A. R. Laddi, and R. R. Akhlolaghi, "Advances in post-harvest management of banana: A review," Scientia Horticulturae, vol. 289, p. 110456, 2022.

[3] R. E. Saragih and A. W. R. Emanuel, "Banana Ripeness Classification Based on Deep Learning using Convolutional Neural Network," in Proc. 2021 3rd East Indonesia Conf. Computer and Information Technology (EIConCIT), Surabaya, 2021, pp. 85–89, doi: 10.1109/EIConCIT50028.2021.9431928.

[4] S. Mutrofin, C. Fatichah, E. Setiawan, and H. Yuniarti, "Convolutional Neural Networks Performance Investigation in Banana Ripeness Classification: Impact of Model, Padding, and Optimizer," in Proc. 2024 9th Int. Conf. Informatics and Computing (ICIC), 2024, doi: 10.1109/ICIC64337.2024.10956746.

[5] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[6] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[7] H. Wu et al., "CvT: Introducing Convolutions to Vision Transformers," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 22–31.

[8] H. Saad, A. P. Ismail, N. Othman, M. H. Jusoh, N. F. Naim, and N. Azam, "Recognizing the Ripeness of Bananas Using Artificial Neural Network Based on Histogram Approach," 2009.

[9] I. A. Sabilla, C. S. Wahyuni, C. Fatichah, and D. Herumurti, "Determining Banana Types and Ripeness from Image using Machine Learning Methods," in Proc. 2019 Int. Conf. Artificial Intelligence and Information Technology (ICAIIT), 2019, pp. 407–412, doi: 10.1109/ICAIIT.2019.8834490.

[10] E. Sunandar, K. D. Hartomo, Y. Nataliani, and I. Sembiring, "Enhanced Banana Ripeness Detection using GLCM and K-NN Methods," 2024.

[11] R. Phoophuangpairoj, T. Ngoenrungrueang, and S. Audomsin, "Ripeness Classification of a Bunch of Bananas Using a CNN," in Proc. 2023 Int. Electrical Engineering Congress (iEECON), Krabi, 2023, pp. 180–183, doi: 10.1109/iEECON56657.2023.10126585.

[12] M. F. Mohamedon, F. Abd Rahman, S. Y. Mohamad, and O. O. Khalifa, "Banana Ripeness Classification Using Computer Vision-based Mobile Application," in Proc. 2021 8th Int. Conf. Computer and Communication Engineering (ICCCE), Kuala Lumpur, 2021, pp. 335–338, doi: 10.1109/ICCCE50029.2021.9467225.

[13] R. G. Baldovino, R. A. U. Lim, P. R. R. Salvador, and E. A. P. Tiamzon, "Real-Time Banana Ripeness Detection and Classification using YOLOv8," in Proc. 2024 9th Int. Conf. Mechatronics Engineering (ICOM), Kuala Lumpur, 2024, pp. 219–223, doi: 10.1109/ICOM61675.2024.10652438.

[14] D. E. Cahya, H. Destika, Z. Cahya, M. N. Alfa, H. Taufiqurrohman, and T. E. Putri, "Ripeness Classification of Cavendish Bananas using Multi-object Detection Approach,"

in Proc. 2023 Int. Conf. Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), 2023, pp. 146–151, doi: 10.1109/ICRAMET60171.2023.10366571.

[15] S. Mutrofin and C. Fatichah, "Automatic Non-Destructive Classification of Banana Ripeness and Varieties Using Digital Imaging: A Systematic Literature Review," in Proc. 2024 9th Int. Conf. Informatics and Computing (ICIC), 2024, doi: 10.1109/ICIC64337.2024.10956818.

[16] S. Thanyacharoen et al., "Deep Learning Approaches to Banana Ripeness Detection," 2025.

[17] R. G. Tiwari, "Transforming Agricultural Diagnosis: Vision Transformer-Based Classification of Banana Leaf Diseases for Precision Farming," in Proc. 2024 1st Int. Conf. Innovative Engineering Sciences and Technological Research (ICIESTR), New Delhi, India, 2024, doi: 10.23919/INDIACom61295.2024.10498201.

[18] B. Xiao, M. Nguyen, and W. Q. Yan, "Fruit Ripeness Identification Using Transformers," Applied Intelligence, 2023, pp. 22488–22499, doi: 10.1007/s10489-023-04799-8.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[20] M. Muhammad and M. Yeasin, "EigenCAM: Class Activation Map using Principal Components," in Proc. Int. Joint Conf. Neural Networks (IJCNN), 2020, doi: 10.1109/IJCNN48605.2020.9206626.

[21] P. B. R. Chirag and S. Bhat, "An Intelligent Approach to Determine Banana Ripeness Stages using Deep Learning Models," 2024.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[23] A. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.

[24] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. Int. Conf. Learning Representations (ICLR), 2019.

[25] R. Agrawal and M. Kumar, "Deep Learning Framework for Banana Shelf Life Classification based on Ripening: BSLC," in Proc. 2023 6th Int. Conf. Information Systems and Computer Networks (ISCON), 2023, pp. 1–5, doi: 10.1109/ISCON57294.2023.10112153.

[26] K. M. S. Callaghan and U. M. Hernandez, "Low-Cost Multisensor Nondestructive Banana Ripeness Estimation Using Machine Learning," IEEE Sensors Journal, vol. 25, no. 5, pp. 8797–8806, Mar. 2025, doi: 10.1109/JSEN.2025.3528250.

[27] T. Shuprajhaa, J. Mathav Raj, P. Suresh Kumar, K. N. Sheeba, and K. Dhayalini, "Deep learning based mobile application for varietal identification and ripeness grading of traditional Indian banana varieties," in Proc. 2023 2nd Int. Conf. Electrical, Electronics,

Information and Communication Technologies (ICEEICT), 2023, doi: 10.1109/ICEEICT56924.2023.10157578.

[28] R. D. C. Magno, P. J. S. Bersamen, and L. D. Valiente Jr., "Monitoring of Banana Ripening Process and Identification of Banana type using FRCNN, LSTM, and Color Segmentation," in Proc. IEEE 14th Int. Conf. Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 2022, doi: 10.1109/HNICEM57413.2022.10109517.

[29] N. Fnu and A. Bansal, "Understanding the Architecture of Vision Transformer and Its Variants: A Review," arXiv preprint arXiv:2401.12842, 2024.

[30] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. Int. Conf. Learning Representations (ICLR), 2015.