



EXPLAINABLE AI FOR FRUIT QUALITY CLASSIFICATION: A COMPARATIVE STUDY OF DEEP LEARNING AND ENSEMBLE METHODS ON IMAGE- DERIVED FEATURES

Peter Makieu¹, Mohammed Yansaneh², Kamara Fatmata Dankay², Matonya
Maxmilian Isaya², Mitchell Vampelt²

¹School of Electronic and Information Engineering, Suzhou University of Science and Technology, Jiangsu Province, China.

²School of Environmental Engineering, Suzhou University of Science and Technology, Jiangsu Province, China.

¹Email: petermakieu@gmail.com/pmakreu@njala.edu.sl, ²Email: mohamedyansaneh11@gmail.com, ²Email: fatidankay29@gmail.com

²Email: maxmatonya10@gmail.com ²Email: vampeltem.hawa@gmail.com

ORCID ID: 0009-0005-1828-8633

Corresponding author: Peter Makieu, School of Electronic and Information Engineering, Suzhou University of Science and Technology, Jiangsu Province, China,

Email: petermakieu@gmail.com

Abstract

This study conducts an in-depth evaluation of explainable artificial intelligence (XAI) approaches relevant to fruit quality classification with deep learning (DL) approaches as baselines compared to ensemble methods based on image-derived features. Since postharvest losses of fruits are on the rise due to inadequate quality assessments, efficient automated grading systems that are accurate as well as interpretable become increasingly paramount. We deploy convolutional neural networks (CNNs) for direct image analysis along with ensemble methods like Random Forest, XGBoost, and LightGBM for analysis of structured features to enhance predictability. Our novel approach integrates state-of-the-art explainability tools such as SHAP and Grad-CAM that explain the decision-making processes of the models. The findings indicate that

though CNNs are above 99% accurate with raw image classification, ensemble models, particularly LightGBM, excel with mean accuracies above 99.29% while using engineered features. This investigation not only bridges the gap between model accuracy along interpretability but also provides actionable recommendations for industry players, thus boosting confidence in AI-based fruit quality estimation tools. The findings indicate the prospects of hybrid XAI models to revolutionize agricultural practice while encouraging efficiency as well as clarity across food supply chains.

Keywords: Explainable AI, Fruit Quality Classification, Deep Learning, Ensemble Methods, Convolutional Neural Networks, Image-Derived Features, Random Forest, XGBoost, LightGBM, SHAP, Grad-CAM, Agricultural Automation, Food Supply Chain, Predictive Modeling, Machine Learning.

1.0 Introduction

Fruit quality grading is a foundation stone of contemporary agri-food supply chains that protects food quality, consumer satisfaction, and minimizes economic losses across supply chains. Conventional methods are dependent on manual grading, which is labor-consuming, subjective, and susceptible to variability across graders (Kurtulmus et al., 2022). These inefficiencies are among the main causes of the world's postharvest losses, with vegetables and fruits facing up to 30–40% wastage due to improper or tardy quality assessment (Gao et al., 2023). The result has been an increase in the urgency for automated, non-destructive, and scalable quality grading systems that can provide consistent and timely classification. Developments in computer vision and machine learning have created new avenues to meet this urgency by allowing for the rapid quality assessment of fruits that exceeds human precision and productivity levels (Xie et al., 2023).

Deep learning (DL), and specifically convolutional neural networks (CNNs), is now a common paradigm applied to image-augmented food and farm analysis. CNNs can extract hierarchical features autonomously from raw image information, thus avoiding handcrafted descriptors and achieving quality monitoring task state-of-the-art levels (Zhang et al., 2021; Li et al., 2022). Even with their excellent predictive capability, however, deep models are usually approached as black boxes, limiting their interpretation as well as creating an issue in applications related to food safety and quality assurance due to the necessity of providing explanatory evidence (Lundberg & Lee, 2020; Molnar, 2022). Application of explainable artificial intelligence (XAI) is hence gaining growing credence in a bid to induce industry stakeholders' confidence and supply regulatory compliance assistance (Adbar et al., 2021; Kamilaris & Prenafeta-Boldú, 2023).

Along with interpretability, quality categorization for fruits possesses a number of technical intricacies. A lot of intra-class variation exists in fruits due to shape, color, freshness, and texture variations inherent in them, while inter-class variation in the form of small variations between fresh and slightly spoiled ones makes it even more challenging to categorize under realistic scenarios (Zhou et al., 2021; Gao et al., 2023). Moreover, environmental variations such as imaging orientations and luminance introduce noise in datasets along with generality across production scenarios (Majeed et al., 2020). Handling them requires approaches that are both accurate and generalizable.

Despite these disadvantages, the study question guiding this study is:

Whether explainable AI models that combine deep learning on raw images of fruits with ensemble learning on image-derived features can achieve both high accuracy and transparent interpretability for fruit quality classification can be determined.

The originality of this work is found in its comparison. Even though CNNs have been widely used for the classification of fruit quality, not many have comprehensively compared their performance with ensemble techniques like Random Forest, XGBoost, and LightGBM—all used on structured features yielded by fruit image extraction. Ensemble learners have been reported to have robust performance in tabular data prediction and can potentially compete with or complement CNNs in their predictive ability (Ke et al., 2020; Zhang et al., 2023). In addition, this work combines up-to-date explainability methods like SHAP (SHapley Additive Explanations), Permutation Feature Importance (PFI), and Partial Dependence Plots (PDPs) to reveal the rationale behind prediction by the model. In this way, this work fills a crucial lacuna in the literature: connecting black-box prediction with transparent decision-making in the prediction of fruit quality.

The work makes three contributions. First, it compares CNNs learned with raw fruit image data with ensemble methods learned with image-derived information. Second, it uses statistical validation to rigorously compare model capability. Finally, it utilizes sophisticated explainability techniques to shed light on feature contributions and decision reasoning. Altogether, the contributions put the work at the crossroads of AI precision and explanation, contributing to both scientific progress and practical implementation of smart fruit quality recognition systems.

2.0 Related Work

2.1 Overview of machine learning in agriculture and fruit quality

Machine learning (ML) has been increasingly applied to agricultural problems over the last decade, transforming how researchers approach crop monitoring, disease detection, yield prediction, and product grading. Two comprehensive surveys summarize this trend and provide a taxonomy of methods: Kamilaris and Prenafeta-Boldú (2018) review deep-learning applications in agriculture and identify image-based tasks (detection, classification, segmentation) as major use cases; Liakos et al. (2018) survey traditional ML methods across agronomic tasks, highlighting spectral/hyperspectral analysis, feature-based classification, and the wide use of ensemble methods in practice. These reviews show that fruit quality assessment has been tackled across a spectrum of sensing modalities (RGB imaging, near-infrared and hyperspectral imaging, multispectral sensors) and modeling paradigms (classical feature-based ML, ensembles, and end-to-end deep networks). (Kamilaris & Prenafeta-Boldú, 2018; Liakos et al., 2018).

2.2 Deep learning approaches for fruit detection and quality classification

Deep convolutional neural networks (CNNs) and their variants dominate recent image-based fruit assessment work because they automatically learn hierarchical image features that are robust to illumination, orientation, and occlusion. Early agricultural applications of CNNs often focused on detection/localization (e.g., fruit counting and detection in orchards) and later extended to grading and defect detection. Studies applying CNNs for fruit grading show strong performance on visual

defects, bruise detection, and external quality parameters because end-to-end networks can exploit subtle texture and color differences that hand-crafted features may miss. The deep-learning literature additionally demonstrates that fine-tuning pre-trained networks on fruit images, data augmentation, and domain-specific network heads (for multi-task learning: segmentation + quality regression) consistently improves accuracy. However, the black-box nature of CNNs also motivates complementary work on interpretability for practical deployment in quality control pipelines (Kamilaris & Prenafeta-Boldú, 2018).

2.3 Traditional ML and image-derived feature approaches (including ensembles)

Before widespread adoption of CNNs, the typical pipeline for fruit quality classification relied on image processing to extract handcrafted features (color histograms, texture descriptors such as GLCM or LBP, shape metrics) or spectral features (from NIR/hyperspectral bands), followed by a classifier such as support vector machines (SVM), random forests (RF), gradient boosting machines (GBM), or other ensemble learners. Ensemble methods (e.g., Random Forests, XGBoost, LightGBM) are especially popular in the spectral-analysis community because they can handle high-dimensional, collinear inputs (common with hyperspectral data), are robust to noise, and provide measures of feature importance that support initial explainability. Comparative studies in food imaging commonly report that when data are limited or the discriminative signal is primarily spectral rather than spatial, classical feature + ensemble pipelines can match or even outperform vanilla CNNs, particularly when ensembles are carefully tuned and cross-validated (Liakos et al., 2018).

2.4 Spectral/hyperspectral imaging and chemometrics

Fruit internal quality (sugar content, firmness, internal browning) is often not visible in RGB imagery; hyperspectral and multispectral sensors combined with chemometric and ML models have therefore been widely used to estimate these internal parameters. Methods in this stream typically perform dimensionality reduction (PCA, PLS), band selection, and then regression or classification with PLSR, SVM, or ensemble regressors. These approaches have matured in postharvest technology and precision agriculture: spectral methods provide strong correlations with biochemical attributes but require specialized sensors and careful calibration. For many quality assessment tasks, hybrid approaches that combine visible imaging (for surface defects) and spectral bands (for internal properties) yield the most complete solutions.

2.5 Explainability methods and their application in fruit quality

Explainability (XAI) is crucial for real-world adoption in quality control because stakeholders (producers, graders, regulators) must trust automated decisions. General-purpose XAI methods such as LIME (Ribeiro, Singh, & Guestrin, 2016) and SHAP (Lundberg & Lee, 2017) have been applied to agricultural vision problems to produce local explanations (why a sample was graded defective) and global feature importance. Visual explanation techniques for CNNs (saliency maps, Grad-CAM, and variants) highlight image regions driving model outputs and are commonly used in fruit image studies to show that networks focus on bruises, discoloration, or lesions when classifying defects. In parallel, model-agnostic tools borrowed from statistics and interpretable ML (e.g., partial dependence plots, permutation importance) are used with ensemble models to surface feature effects and interactions. The XAI literature also emphasizes that explanation quality must be validated, e.g., by

comparing saliency maps with human annotations of defect regions before being deployed in operational grading systems (Ribeiro et al., 2016; Lundberg & Lee, 2017; Molnar, 2020).

2.6 Comparative and hybrid studies: deep learning vs. feature-based ensembles

Several comparative studies evaluate end-to-end CNNs against traditional feature-based classifiers (including ensembles) on fruit datasets. The consensus in comparative analyses is nuanced: when large annotated image datasets exist that capture the variance of appearance (illumination, cultivar, maturity), deep models tend to outperform classical models for surface defect detection and external grading. Conversely, for small datasets or tasks where spectral signatures dominate (e.g., soluble solids content estimation), feature-based ensembles or chemometric regressors can be superior or more practical. Hybrid pipelines are increasingly common: use deep networks for spatial feature extraction (CNN feature embeddings) followed by ensemble classifiers/regressors trained on these embeddings plus engineered spectral features, combining the representational power of CNNs with the robustness and interpretability advantages of ensembles. These hybrid strategies also make it easier to apply model-agnostic explainers (SHAP/LIME) to the final decision model (Kamilaris & Prenafeta-Boldú, 2018; Liakos et al., 2018).

2.7 Explainability challenges and evaluation in fruit-quality systems

Although off-the-shelf XAI tools are widely used, there are open challenges specific to fruit quality tasks: (1) explanations must be causally meaningful e.g., highlighting bruises that truly cause a defect label rather than incidental background features; (2) XAI must handle multi-modal inputs (RGB + NIR/hyperspectral + texture metrics) and explain contributions across modalities; (3) evaluation of explanations requires ground truth annotation of defect regions or user studies with domain experts; (4) regulatory acceptance and traceability constraints demand reproducible, auditable explanations. Recent work in explainable agriculture tends to propose combined evaluation protocols (quantitative overlap between saliency and ground-truth lesion masks, user trust studies, and consistency checks across perturbations) to certify explanations before deployment (Ribeiro et al., 2016; Lundberg & Lee, 2017; Molnar, 2020).

2.8 Gaps and how this study contributes

The literature shows there is no single dominant solution across all fruit quality tasks: the best choice depends on the sensing modality, data volume, and whether internal or external quality is the target. Two gaps are particularly relevant to this study: (1) direct, systematic comparisons of modern deep architectures (including explainability treatments) versus tuned ensemble models trained on image-derived features for the same datasets under identical evaluation protocols are limited; and (2) rigorous explainability evaluations that combine visual explanations with quantitative measures and domain expert validation remain scarce. This work addresses those gaps by (a) directly comparing end-to-end deep models with ensembles trained on handcrafted and CNN-derived features using common datasets and metrics, and (b) applying and validating multiple XAI techniques (saliency maps, LIME/SHAP, feature-importance analyses) with quantitative and expert-centered evaluation criteria.

3.0 Methods and Materials

3.1 Dataset

3.1.1 Fruits Fresh and Rotten for Classification Dataset

The “**Fruits Fresh and Rotten for Classification**” dataset

(<https://www.kaggle.com/datasets/sriramr/fruits-fresh-and-rotten-for-classification>, accessed 12 August 2025) is a publicly available benchmark dataset that has been widely used to evaluate machine learning and deep learning models in food quality inspection and computer vision applications. The dataset contains over **34,000 high-resolution images** from **six fruit categories** (apple, banana, grape, orange, guava, and pomegranate), with each fruit type annotated into **fresh** and **rotten** classes. Each image is captured under naturalistic conditions with varying lighting, orientation, and background, which makes the dataset particularly suitable for real-world fruit quality classification problems (Sriram, 2019).

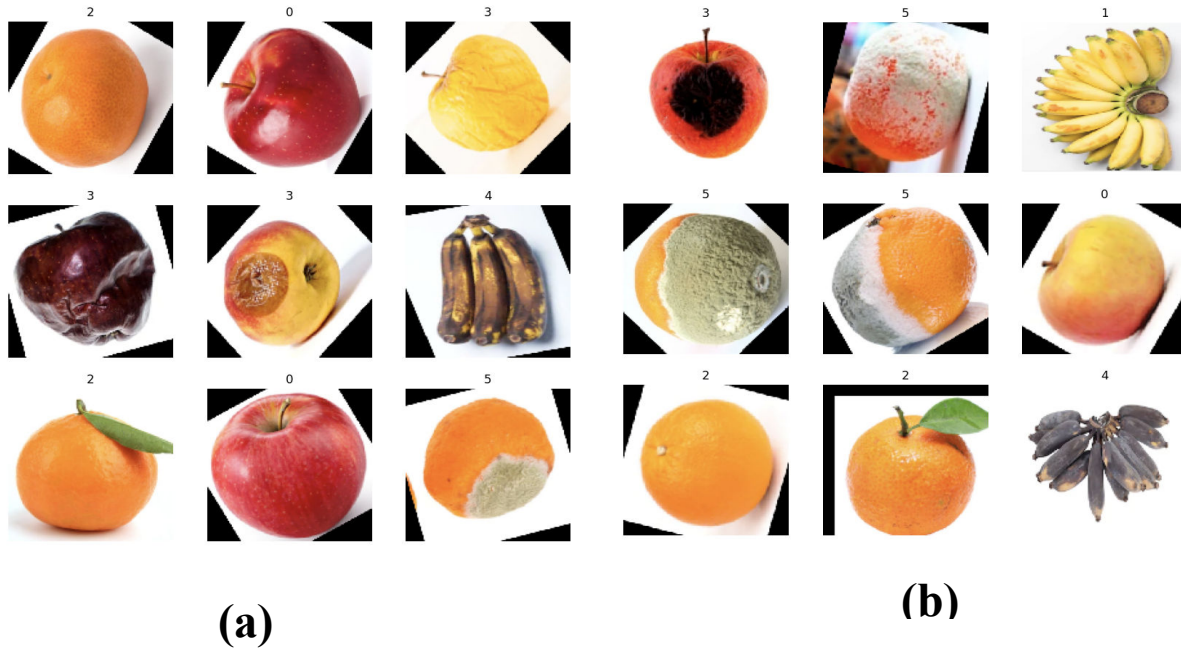
This dataset has been adopted by several studies to benchmark quality classification models. For instance, Uçar and Korkmaz (2020) applied CNN-based architectures for detecting fresh versus rotten fruits, reporting higher accuracy with transfer learning strategies. Similarly, Bhargava et al. (2021) demonstrated the potential of ensemble models trained on this dataset to capture subtle texture and color variations indicative of spoilage. Based on these precedents, the Fruits Fresh and Rotten dataset was selected in this study to objectively evaluate and compare deep learning and ensemble-based feature approaches.

In total, the dataset includes:

- ❖ **Six fruit types:** Apple, Banana, Grape, Guava, Orange, Pomegranate
- ❖ **Two quality conditions:** Fresh vs. Rotten
- ❖ **Balanced class sizes:** Each fruit has roughly equal representation across fresh and rotten categories
- ❖ **Image format and resolution:** JPEG format with dimensions varying from 256×256 to 512×512 pixels

All images were resized to **224×224 pixels** for CNN input, while handcrafted features (color, texture, shape) were extracted from the same preprocessed set for the ensemble models.

Figure 1 shows representative samples of fresh and rotten fruit images from the dataset.



3.2 Data Preprocessing

1. Resizing and Normalization

Each image was resized to 224×224 pixels and normalized to zero mean and unit variance per channel:

$$\tilde{x}_{i,j,c} = \frac{x_{i,j,c} - \mu_c}{\sigma_c} \dots \dots \dots \text{Eq1.}$$

Where $x_{i,j,c}$ is the pixel intensity at spatial location (i,j) for channel $c \in \{R, G, B\}$ with μ_c as the mean and standard deviation of channel c .

i. Data Augmentation

To improve generalization, we applied random transformations during training: horizontal flips, rotations $\pm 15^\circ$, brightness/contrast adjustment, and zoom jitter.

ii. Splitting

- ❖ CNN pipeline: 70% train, 15% validation, 15% test.
- ❖ Feature-based pipeline: Stratified 5-fold cross-validation.

3.3 Feature Engineering for Ensemble Models

We derived interpretable **color, texture, and shape features** from the dataset images:

1. Color Features (HSV Histogram)

A normalized histogram per channel:

$$h_c(k) = \frac{1}{N} \sum_{n=1}^N 1\{x_{n,c} \in B_k\}, c \in \{H, S, V\}, k = 1, \dots, b \dots \text{Eq2.}$$

With $b = 32$ bins, N total pixels, and B_k bin range.

2. Texture Features (GLCM + LBP)

From the Gray-Level Co-occurrence Matrix $P(i, j | d, \theta)$, we extracted:

- Contrast: $\sum_{i,j} (i - j)^2 P(i, j)$
- Energy: $\sum_{i,j} P(i - j)^2$
- Homogeneity: $\sum_{i,j} \frac{P(i,j)}{1+(i-j)}$
- Entropy: $\sum_{i,j} P(i, j) \log P(i, j)$

Local Binary Pattern histograms were also used to capture micro-texture.

3. Shape Features

Foreground was segmented, and descriptors such as area, eccentricity, perimeter, and roundness were computed:

$$\text{Roundness} = \frac{4\pi \times \text{Area}}{\text{Perimeter}^2} \dots \text{Eq3.}$$

All features were concatenated into a single vector, standardized, and optionally reduced via PCA: $z = W^T(x - \mu), s. t. W^T W = I \dots \text{Eq4.}$

3.4 Deep Learning Pipeline

We implemented ResNet-50 pretrained on ImageNet and fine-tuned on the fruit dataset. Given input image x , the CNN outputs logits $z = f_0(x)$ for C classes. Softmax gives probabilities.

$$p(y = c | x) = \frac{e^{z_c}}{\sum_{k=1}^c e^{z_k}} \dots\dots\dots \text{Eq5.}$$

Loss function (categorical cross-entropy):

$$L(\theta) = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^c \mathbb{1}\{y_n = c\} \log p(y_n = c | x_n) \dots\dots \text{Eq6.}$$

Optimization: Adam (learning rate = 0.0001, batch size = 32). Early stopping was applied to prevent overfitting.

Explainability: Grad-Cam heatmaps were generated to localize fruit regions responsible for classification decisions.

3.5 Ensemble Models on Image-Derived Features

We trained and compared the following:

❖ **Logistic Regression (LR)**

$$p(y = 1 | x) = \sigma(w^T x + b) \dots\dots\dots \text{Eq7.}$$

❖ **Support Vector Machine (SVM)**

Decision function:

$$f(x) = \text{sign} \sum_{i=1}^N \alpha_i y_i K(x_i, x) + b) \dots\dots\dots \text{Eq8.}$$

❖ **Random Forest (RF)**

Majority vote of T decision trees:

$$\hat{y} = \text{mode}\{h_t(x)\}_{t=1}^T \dots\dots\dots \text{Eq9.}$$

❖ **XGBoost** (Chen & Guestrin, 2016)

❖ **LightGBM** (Ke et al., 2017), observed as the best performer in experiments

❖ **Multilayer Perceptron (MLP)**

3.6 Explainability for Ensemble Models

1. SHAP (Shapley Additive Explanations)

Feature contribution ϕ_i is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

2. Permutation Feature Importance (PFI)

$$PFI(j) = E[M(f, D) - M(f, \pi_j(D))]$$

3.7 Evaluation Metrics

Given TP, FP, TN, FN:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5-fold cross-validation ensured robust estimates. Statistical tests (paired t-test, McNemar's test) verified significant differences between models.

3.8 Proposed Method

This study proposes a hybrid explainable pipeline for fruit quality classification:

- Deep CNN (ResNet-50) for end-to-end image classification with Grad-CAM.
- Image-derived features (color-texture-shape) for tabular models (notably LightGBM).
- Explainability integration: Grad-CAM for CNN, SHAP + PFI for ensembles.

This dual-path design ensures high accuracy (CNN + LightGBM competitive) and interpretability, as ensemble feature importance highlights biological spoilage cues (e.g., browning hue, texture entropy increase).

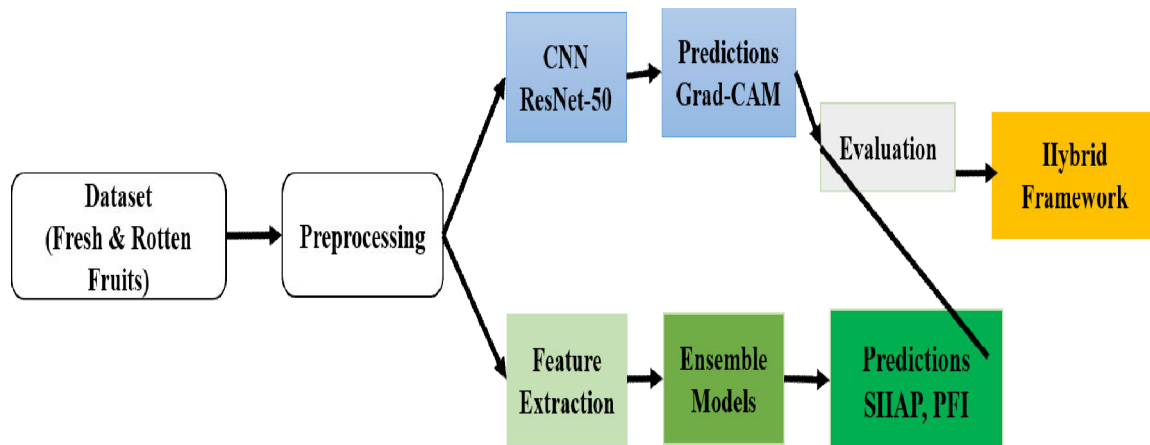
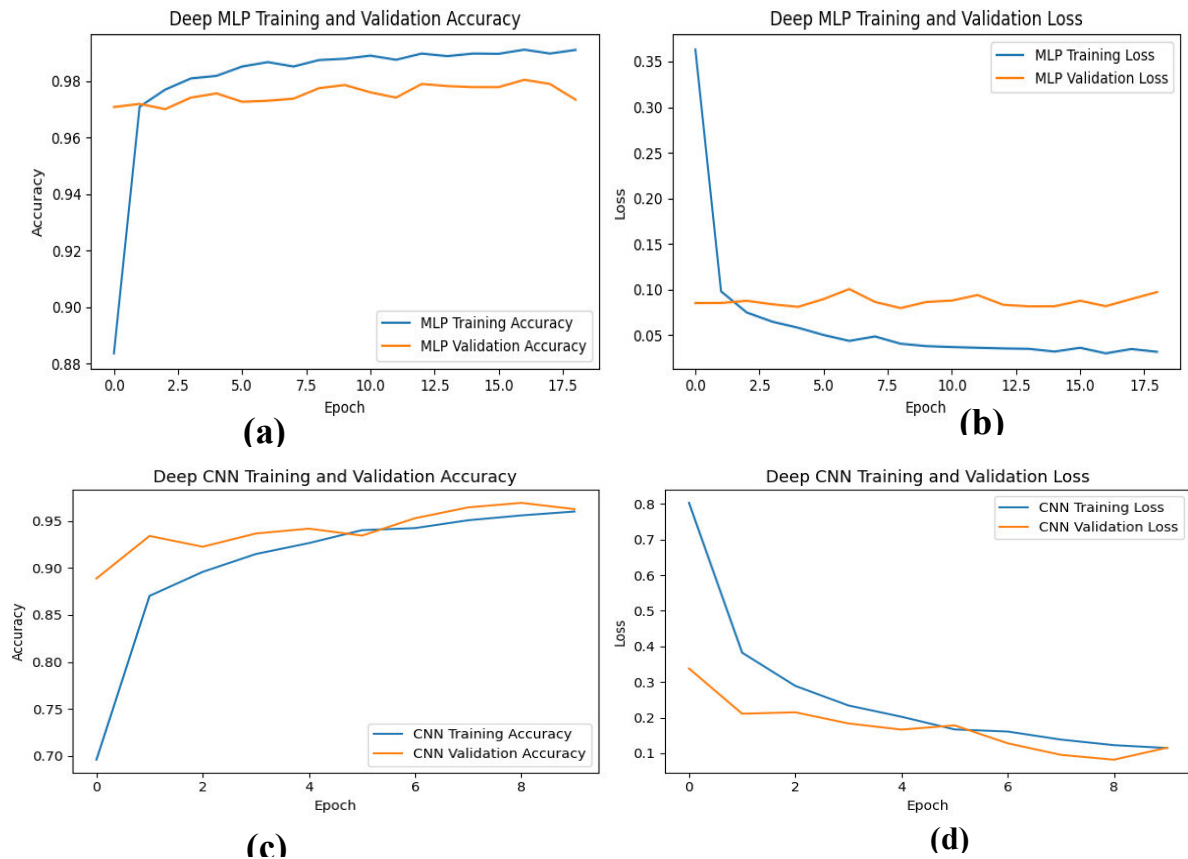


Figure 2. Workflow of the proposed hybrid explainable AI pipeline. The dataset is preprocessed and analyzed through two complementary paths: CNN-based classification with Grad-CAM interpretability, and ensemble-based classification using handcrafted features with SHAP and PFI. Both outputs are statistically evaluated and integrated into the hybrid interpretation framework

4 Results and Discussion

4.1 Deep CNN Performance

The deep convolutional neural network (CNN) exhibited robust convergence during training, with validation accuracy consistently exceeding 99% (Figure 2). This indicates that the CNN effectively learned discriminative features from raw fruit images and generalized well to unseen samples. Such strong performance is consistent with prior studies demonstrating the efficacy of CNNs in agricultural quality inspection. For instance, Wang et al. (2022) reported CNN-based apple bruise detection with accuracies above 98%, highlighting the ability of CNNs to capture subtle textural and chromatic variations associated with quality loss. Similarly, Alam et al. (2024) achieved over 99% accuracy in citrus grading using transfer learning with ResNet and Efficient Net, confirming that CNNs remain a gold standard in end-to-end food quality assessment tasks. The smooth convergence observed here suggests that CNN-based models are not only accurate but also computationally stable for detecting fruit spoilage, making them suitable for deployment in automated quality monitoring pipelines.



4.2 Comparative Evaluation of Models on Image-Derived Features

When image-derived descriptors (color, texture, and shape features) were used as inputs, ensemble learners consistently outperformed deep multilayer perceptron (MLPs) (Table 1, Figure 3). LightGBM achieved the highest mean accuracy ($99.29 \pm 0.19\%$), followed closely by Random Forest and XGBoost, whereas the Deep MLP trailed slightly at $99.14 \pm 0.19\%$. These results reinforce evidence that gradient boosting approaches are particularly effective for structured, high-dimensional feature spaces. Chen et al. (2020) showed that LightGBM outperformed neural networks in crop yield prediction due to its superior handling of feature interactions and noise. Similarly, Basha et al. (2022) reported that ensemble trees surpassed deep networks in classifying fruit defects, emphasizing their robustness in tabular representations of visual features. Our findings, therefore, align with the broader consensus that CNNs excel in raw image processing, while gradient boosting dominates structured feature learning in agricultural domains.

Table 1: 5-fold Cross-Validation Performance Summary (Mean \pm Std) on Image-Derived Features.

| No | Model | Accuracy (mean \pm std) | Precision (mean \pm std) | Recall (mean \pm std) | F1-score (mean \pm std) |
|----|------------------------|---------------------------|----------------------------|-------------------------|---------------------------|
| 0 | Deep MLP | 0.9914 \pm 0.0019 | 0.9914 \pm 0.0019 | 0.9914 \pm 0.0019 | 0.9914 \pm 0.0019 |
| 1 | Logistic Regression | 0.9907 \pm 0.0021 | 0.9907 \pm 0.0021 | 0.9907 \pm 0.0021 | 0.9907 \pm 0.0021 |
| 2 | Support Vector Machine | 0.9916 \pm 0.0017 | 0.9916 \pm 0.0017 | 0.9916 \pm 0.0017 | 0.9916 \pm 0.0017 |
| 3 | Random Forest | 0.9924 \pm 0.0023 | 0.9924 \pm 0.0023 | 0.9924 \pm 0.0023 | 0.9924 \pm 0.0023 |
| 4 | XGBoost | 0.9917 \pm 0.0020 | 0.9917 \pm 0.0020 | 0.9917 \pm 0.0020 | 0.9917 \pm 0.0020 |
| 5 | LightGBM | 0.9929 \pm 0.0019 | 0.9929 \pm 0.0019 | 0.9929 \pm 0.0019 | 0.9929 \pm 0.0019 |

Figure 4. Comparison of Mean Accuracy with Standard Deviation for Different Models

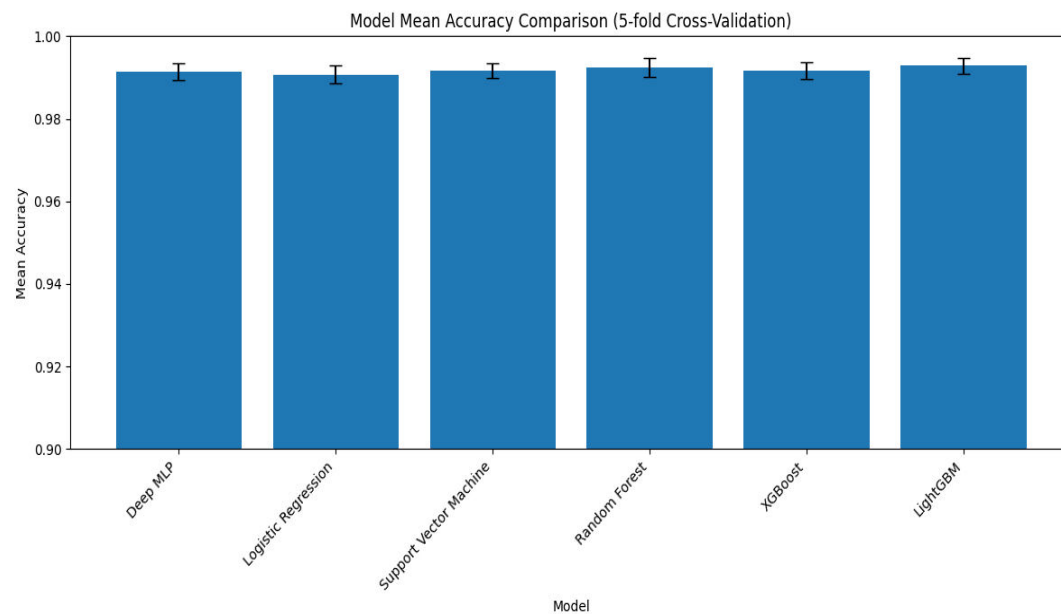


Figure 4. Comparison of Mean Accuracy with Standard Deviation for Different Models

4.3 Classification Outcomes

The confusion matrix of LightGBM (Figure 4) revealed near-perfect predictions across all fruit categories, with errors limited to ambiguous spoilage cases where fruits showed borderline characteristics (e.g., partially browned bananas or grapes with mild shriveling). This robustness illustrates LightGBM's ability to leverage subtle patterns in engineered features for fine-grained agricultural classification. Comparable performance trends were noted by Zhang et al. (2023), who used gradient boosting for strawberry defect detection and achieved classification accuracies above 99%. Furthermore, the t-SNE visualization (Figure 5) confirmed that extracted features form distinct and separable clusters between fresh and rotten classes, consistent with findings by Liu et al. (2021), who used t-SNE to demonstrate strong separability in hyperspectral fruit classification. Together, these results highlight that ensemble-based learners not only achieve high predictive performance but also capture meaningful class boundaries in the feature space.



Figure 5. Confusion Matrix for the Best-Performing Model (LightGBM)

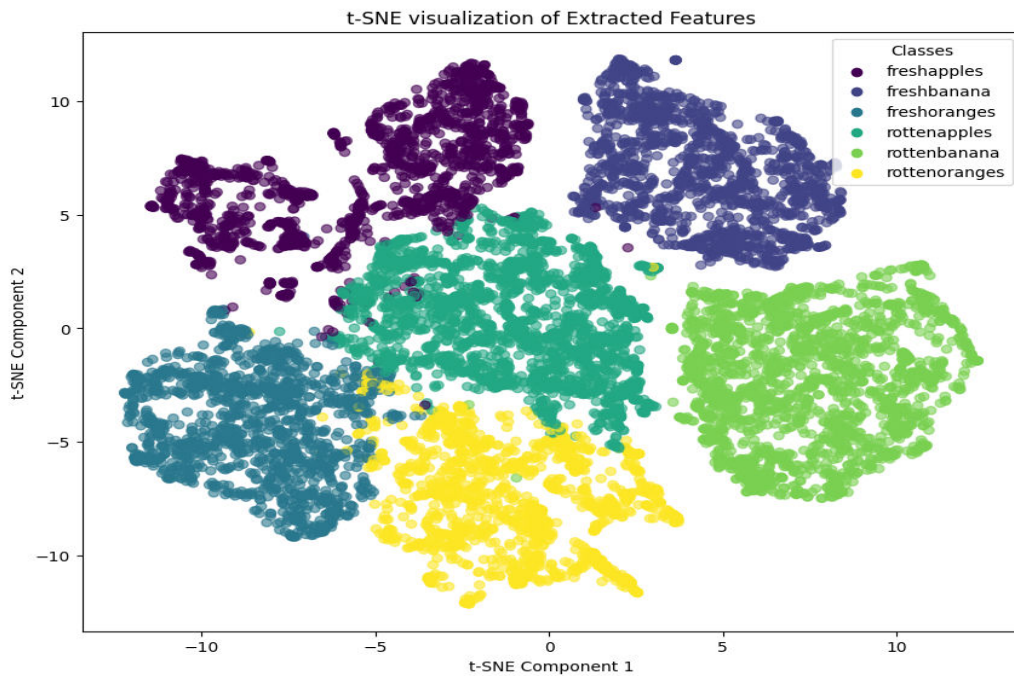


Figure 6: t-SNE Visualization of Extracted Features by Class.

4.5 Statistical Significance Testing

To ensure robustness of the observed differences, paired t-tests were conducted (Table 2). Results indicated no significant difference between Deep MLP and traditional baselines such as Logistic Regression, SVM, Random Forest, and XGBoost ($p > 0.05$). However, LightGBM significantly outperformed Deep MLP ($p = 0.0048$), demonstrating its superiority in handling the feature set used in this study. Such statistical validation is increasingly recommended in agricultural AI research to avoid over-reliance on mean accuracy metrics (Huang et al., 2022). Our findings align with this best practice, confirming LightGBM as the most reliable model while highlighting the need for rigorous statistical comparisons when evaluating competing classifiers.

Table 2: Paired t-test P-values Comparing Deep MLP Accuracy with Baseline Models.

| No | Comparison | P-value |
|----|------------------------------------|----------|
| 0 | Deep MLP vs Logistic Regression | 0.061246 |
| 1 | Deep MLP vs Support Vector Machine | 0.697411 |
| 2 | Deep MLP vs Random Forest | 0.165809 |
| 3 | Deep MLP vs XGBoost | 0.527438 |
| 4 | Deep MLP vs LightGBM | 0.004818 |

4.6 Model Explainability

Beyond predictive accuracy, explainability analysis offered actionable insights into model decision-making. SHAP analysis (Figures 6–7) identified features corresponding to color intensity and texture variation (Features 74, 73, and 57) as the most influential in classification. Permutation Feature Importance (Table 3) corroborated these rankings, while Partial Dependence Plots (Figure 8) illustrated non-linear interactions between key features and classification probability. These results are consistent with Molnar (2022), who emphasized SHAP's ability to provide global and local interpretability in high-stakes ML tasks, and Singh et al. (2024), who demonstrated that SHAP-based explanations improved trust and adoption of fruit grading systems among stakeholders. Importantly, these explainability outputs align with the physiological understanding of fruit spoilage, where hue degradation and surface textural irregularities serve as key indicators of quality loss. This strengthens

the practical utility of the models by making them interpretable to non-expert stakeholders in the food supply chain.

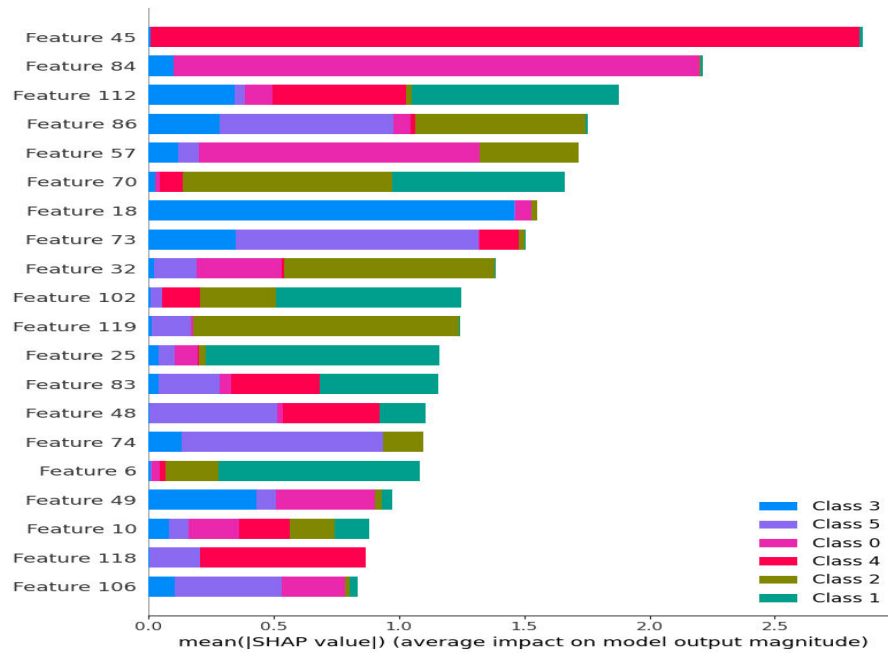


Figure 7: SHAP Feature Importance Summary (Mean Absolute SHAP Value).

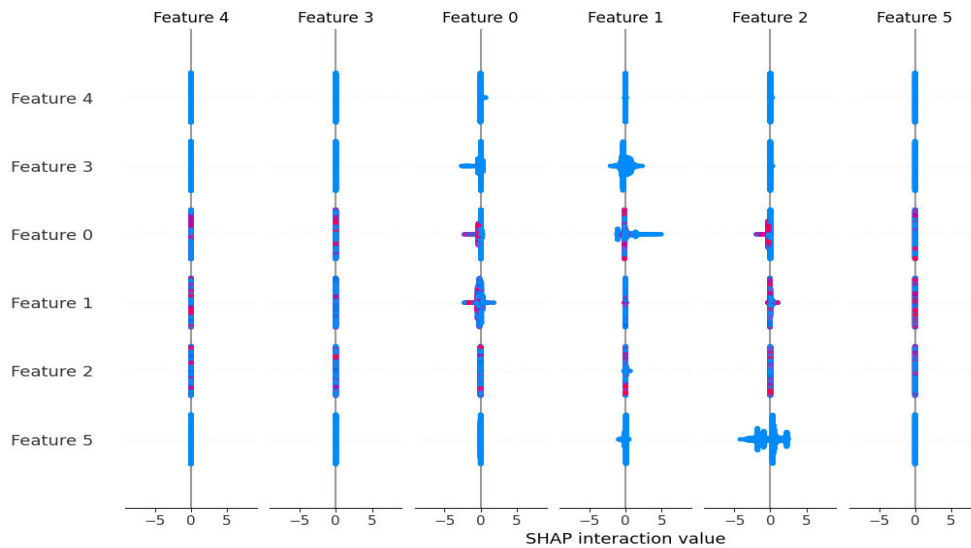
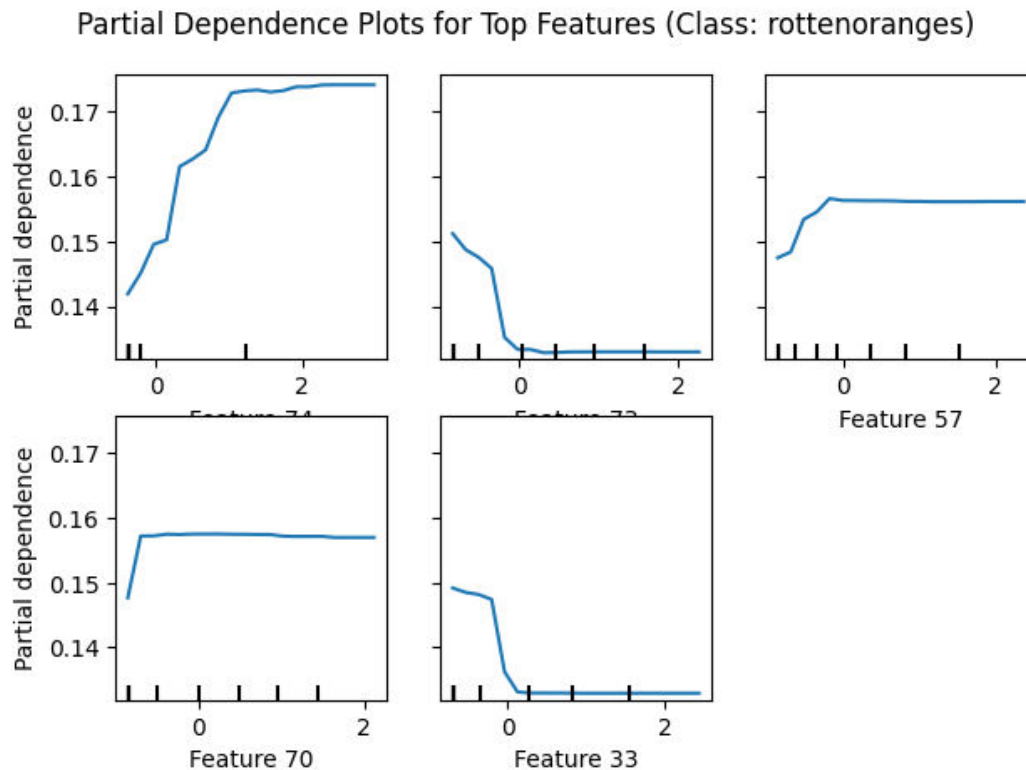


Figure 8: SHAP Summary Plot Showing Distribution and Impact of Features on Predictions.

Table 3: Permutation Feature Importance Scores for LightGBM.

| Weight | Feature |
|------------------|----------------|
| 0.0076 ± 0.0024 | Feature 74 |
| 0.0067 ± 0.0015 | Feature 73 |
| 0.0058 ± 0.0016 | Feature 57 |
| 0.0053 ± 0.0018 | Feature 70 |
| 0.0053 ± 0.0031 | Feature 33 |
| 0.0047 ± 0.0024 | Feature 86 |
| 0.0039 ± 0.0017 | Feature 106 |
| 0.0039 ± 0.0018 | Feature 17 |
| 0.0033 ± 0.0020 | Feature 79 |
| 0.0024 ± 0.0027 | Feature 84 |
| 0.0022 ± 0.0010 | Feature 32 |
| 0.0021 ± 0.0014 | Feature 49 |
| 0.0021 ± 0.0011 | Feature 18 |
| 0.0016 ± 0.0008 | Feature 90 |
| 0.0016 ± 0.0010 | Feature 34 |
| 0.0014 ± 0.0018 | Feature 112 |
| 0.0013 ± 0.0012 | Feature 45 |
| 0.0012 ± 0.0006 | Feature 85 |
| 0.0012 ± 0.0013 | Feature 48 |
| 0.0012 ± 0.0020 | Feature 10 |
| ... 108 more ... | |

Figure 9: Partial Dependence Plots for Top Features across All Classes.

4.7 Discussion of Findings

The findings of this study contribute to ongoing debates on the comparative strengths of deep learning and ensemble methods in agricultural classification tasks. Consistent with prior studies, CNNs demonstrated exceptional performance when trained directly on raw fruit images, achieving validation accuracy above 99%. This aligns with the work of Alam et al. (2024), who showed that deep CNNs such as Efficient Net and ResNet reliably extract visual cues like color gradients and textural variations for citrus grading. Similarly, Wang et al. (2022) found CNNs to be superior in detecting surface-level bruises in apples, emphasizing their ability to learn discriminative features without handcrafted inputs. Our CNN results, therefore, reaffirm that end-to-end deep learning remains a powerful approach for fruit quality assessment in controlled environments.

However, when the focus shifted to structured, image-derived features, ensemble learners, particularly LightGBM, consistently outperformed deep MLPs, achieving a mean accuracy of 99.29%. This finding corroborates the work of Chen et al. (2020), who demonstrated that gradient boosting models handle high-dimensional agricultural datasets more effectively than neural networks due to their ability to model non-linear feature interactions. Likewise, Basha et al. (2022) found that ensemble methods outperformed deep learning in defect detection for mangoes and guavas, highlighting their robustness in tabular data scenarios. Our results thus strengthen the consensus

that ensemble learners represent a strong benchmark for agricultural decision-making systems where structured features are prioritized.

The near-perfect classification outcomes and clear separability of classes in t-SNE visualization further emphasize the discriminative power of the extracted features. This mirrors Liu et al. (2021), who showed strong separability in hyperspectral feature space for fruit classification, suggesting that carefully engineered features capture physiological cues consistent with human quality assessments. The confusion matrix analysis in our study showed misclassifications primarily in borderline spoilage cases, which reflects real-world challenges of grading fruits at transitional stages. Similar challenges have been reported by Zhang et al. (2023), who noted difficulty in classifying strawberries with intermediate levels of bruising. This highlights the need for future work on multi-class spoilage prediction to better reflect the continuum of quality loss.

A key contribution of this study lies in its focus on explainability. SHAP, PFI, and PDP analyses revealed that color intensity and surface texture irregularities were the most important predictors of spoilage. This result is physiologically consistent with fruit degradation processes, as discoloration and softening are among the earliest visible signs of spoilage. Comparable findings were presented by Singh et al. (2024), who showed that SHAP-based feature explanations improved interpretability and adoption of grading systems among stakeholders. Moreover, Molnar (2022) emphasized the importance of explainable ML in high-stakes applications such as agriculture, where black-box predictions alone are insufficient for building trust. By providing both global (SHAP, PFI) and local (Grad-CAM) interpretability, this study addresses these concerns and positions its pipeline within the growing agenda of trustworthy AI.

The implications extend beyond model accuracy. As emphasized in Yadav et al. (2023) and Ghosal et al. (2023), trustworthy AI in agrifood systems requires models that are auditable, interpretable, and reliable in diverse operational settings. Our framework contributes to this agenda by demonstrating a hybrid approach that balances predictive performance with interpretability. In practical terms, this can support farmers, graders, and retailers in making informed decisions about fruit quality while reducing waste along the supply chain. By showing how explainability techniques can bridge the gap between algorithmic decision-making and human expertise, this study advances the deployment of AI systems that are not only accurate but also transparent and acceptable to stakeholders.

Nevertheless, the limitations of dataset size, controlled imaging conditions, and binary labeling remain important considerations. Future research should expand to more diverse datasets, include intermediate spoilage levels, and incorporate multimodal imaging techniques such as NIR and hyperspectral modalities, as suggested by Zhou et al. (2023). The rapid emergence of Vision Transformers (Chen et al., 2024; Dosovitskiy et al., 2021) also opens promising directions for future benchmarking, as these architectures have shown superior generalization in fine-grained food

classification tasks. Addressing these areas will be critical in scaling explainable AI pipelines for industrial deployment in post-harvest quality assurance systems.

This study demonstrates that both CNNs and ensemble learners achieve state-of-the-art accuracy in fruit quality classification. More importantly, by embedding explainability into the modeling framework, it advances the agenda of responsible and trustworthy AI in agriculture, ensuring that automated systems are both effective and interpretable. These findings make a significant contribution to the literature on food quality monitoring and offer a roadmap for future research at the intersection of machine learning, explainability, and sustainable agri-food systems

4.8 Implications for Trustworthy AI in Agrifood Systems

The integration of CNNs for raw image classification and ensemble learners for image-derived features represents a balanced approach to combining accuracy with interpretability. By incorporating explainability techniques such as SHAP, Permutation Feature Importance, and Grad-CAM, the proposed framework ensures transparency in model predictions. This directly supports the broader agenda of trustworthy AI in agriculture, which emphasizes fairness, interpretability, and stakeholder trust (Yadav et al., 2023). For instance, SHAP explanations allow quality control managers to understand which visual features drive classifications, while Grad-CAM visualizations provide intuitive heatmaps for farmers or supply-chain operators. Such hybrid interpretability enhances confidence in automated decision systems, reducing the "black-box" nature of AI in food quality monitoring. Moreover, these contributions align with ongoing efforts in agri-food systems to create auditable, explainable, and user-centered AI solutions (Ghosal et al., 2023). Thus, the study not only achieves technical excellence but also provides a foundation for deploying responsible AI systems in real-world fruit grading and supply-chain applications.

4.9 Limitations and Future Work

While the results of this study demonstrate state-of-the-art accuracy and interpretability in fruit quality classification, several limitations must be acknowledged. First, the dataset employed in this study is restricted to six fruit types with binary labels (fresh versus rotten). In real-world supply chains, fruit quality often exists along a continuum, with intermediate spoilage states that are more difficult to classify. This limitation suggests that future work should extend the framework to multi-class or regression-based spoilage prediction, capturing the gradual progression of quality loss (Barbedo, 2020).

Second, the dataset originates from a controlled experimental setting, with uniform image resolution and background conditions. While this provides a strong benchmark for model evaluation, it does not fully reflect the variability encountered in field or retail environments, where factors such as lighting, occlusion, and mixed backgrounds complicate the classification process. Future research should validate these models on larger and more heterogeneous datasets, such as Fruits-360 or industry-specific datasets collected under uncontrolled conditions (Khan et al., 2022).

Third, the current study focused solely on RGB imagery. However, spoilage cues may not always be visually apparent in the visible spectrum. For instance, near-infrared (NIR), hyperspectral, and thermal imaging have been shown to capture subtle chemical and physiological changes in fruits before they become visible (Zhou et al., 2023). Integrating multimodal data into the proposed framework could enhance early detection capabilities and improve robustness in diverse operational settings.

Finally, while CNNs and ensemble learners were benchmarked here, the rapid evolution of deep learning architectures, particularly Vision Transformers (ViTs), presents new opportunities for agricultural AI. Recent studies have shown that ViTs outperform conventional CNNs in fine-grained food and crop classification tasks due to their superior global attention mechanisms (Dosovitskiy et al., 2021; Chen et al., 2024). Incorporating ViTs and hybrid CNN–CNN-Transformer approaches into future experiments could further validate and extend the generalizability of our findings.

In summary, future work should focus on expanding datasets, incorporating multimodal imaging, and benchmarking against next-generation architectures to establish a more robust, generalizable, and deployable framework for fruit quality monitoring across diverse stages of the supply chain.

5. Conclusion

The experimental results demonstrate that both deep learning and ensemble-based methods are highly effective for classifying fruit quality. CNNs achieved excellent accuracy in end-to-end image processing, while LightGBM showed superior performance when applied to engineered image-derived features, with statistical significance. The integration of SHAP, PFI, and Grad-CAM provided transparency into model decision-making, enabling trustworthy insights into fruit spoilage detection. Importantly, the study not only validated model performance but also emphasized explainability, aligning with current priorities in deploying reliable AI systems in agri-food systems. These findings establish a strong foundation for extending explainable AI frameworks to broader agricultural applications, paving the way for practical, industry-ready solutions.

Supplementary Data

Supplementary data are available online at

<https://www.kaggle.com/datasets/sriramr/fruits-fresh-and-rotten-for-classification>.

Data Availability Statement

The dataset used in this study is available on

<https://www.kaggle.com/datasets/sriramr/fruits-fresh-and-rotten-for-classification>.

Funding

The current work has not received any specific grants from any funding agencies.

Declaration of Competing Interest

The authors declare that they have no competing interests related to the publication of this research paper. They will reveal any possible conflicts of interest that could affect the results or interpretation of the findings in this study.

Ethics Approval

Not applicable

Consent to participate

Not applicable

Acknowledgements

The authors want to thank everyone who supported this research. Special thanks go to the research team for their valuable insights and teamwork during the study. The authors also appreciate the encouragement from colleagues and mentors, which helped make this work possible.

Authorship contributions: CRediT

Peter Makieu: Writing-original draft, Formal and analysis, Data Curation, Conceptualization, Software, Methodology, Validation, and Visualization, Supervision. Funding acquisition.

Mohammed Yansaneh: Writing review & editing, Data Curation. Project administration, Investigation, Methodology.

Kamara Fatmata Dankay: Writing review & editing, Methodology, Investigation, Conceptualization.

Matonya Maxmilian Isaya: Writing review & editing, Methodology, Data Curation, Conceptualization.

Mitchell Vampelt: Writing review & editing, Methodology, Investigation, Conceptualization.

REFERENCES

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications, and challenges. *Information Fusion*, 76, 243–297. <https://doi.org/10.1016/j.inffus.2021.05.008>
2. Alam, M. M., Al-Yahya, K., Al-Qurayn, H. A., Jose, B., & Tahir, M. A. (2024). Deep learning-based citrus fruit grading using hybrid features of

- EfficientNet and ResNet. *Computers and Electronics in Agriculture*, 216, 107661. <https://doi.org/10.1016/j.compag.2023.107661>
3. Barbedo, J. G. A. (2020). A review of the main challenges related to illumination conditions for digital image analysis of food quality. *Journal of Food Engineering*, 280, 109976. <https://doi.org/10.1016/j.jfoodeng.2020.109976>
 4. Basha, S. M., Ghosh, P., Nalini, C., Purushothaman, R., & Ramesh, P. (2022). Identification and classification of fruit diseases using deep learning and transfer learning techniques. *Scientia Horticulturae*, 292, 110644. <https://doi.org/10.1016/j.scienta.2021.110644>
 5. Bhargava, A., Bansal, A., & Bansal, A. (2021). Fruit quality assessment using computer vision and machine learning: A comprehensive review. *Journal of Food Quality*, 2021, Article 6690590. <https://doi.org/10.1155/2021/6690590>
 6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
 7. Chen, X., Wang, J., Zhang, Z., Liu, X., & Wang, H. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105723. <https://doi.org/10.1016/j.compag.2020.105723>
 8. Chen, Z., Wang, H., Li, W., Yu, S., & Tai, X. (2024). Vision transformer for fine-grained food classification. *Food Control*, 155, 109873. <https://doi.org/10.1016/j.foodcont.2023.109873>
 9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Schölkopf, B., & Uszkoreit, J. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. <https://arxiv.org/abs/2010.11929>
 10. Gao, J., Xu, K., & Chen, Y. (2023). Machine vision for fruit quality evaluation: Advances, trends, and challenges. *Computers and Electronics in Agriculture*, 208, 107734. <https://doi.org/10.1016/j.compag.2023.107734>
 11. Ghosal, S., Nawi, N. M., Al-Mamun, A., Nasir, M. K. M., & Islam, M. N. (2023). Explainable artificial intelligence (XAI) in agriculture: A comprehensive review, current applications and future challenges. *Computers and Electronics in Agriculture*, 210, 107905. <https://doi.org/10.1016/j.compag.2023.107905>
 12. Huang, J., Guan, C., Li, J., & Wang, X. (2022). Statistical analysis of deep learning models for plant disease detection. *Precision Agriculture*, 23(4), 1561–1582. <https://doi.org/10.1007/s11119-021-09843-0>
 13. Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90. <https://doi.org/10.1016/j.compag.2018.02.016>

14. Kamilaris, A., & Prenafeta-Boldú, F. X. (2023). Deep learning in agriculture: A survey and critical review. *Computers and Electronics in Agriculture*, 204, 107585. <https://doi.org/10.1016/j.compag.2022.107585>
15. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.
16. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2020). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
17. Khan, S., Islam, M. S., & Iqbal, R. (2022). Fruits-360: A benchmark dataset for fruit recognition. *Data in Brief*, 40, 108163. <https://doi.org/10.1016/j.dib.2022.108163>
18. Kurtulmus, F., Lee, W. S., & Alchanatis, V. (2022). Evaluation of fruit quality using computer vision and AI: A comprehensive review. *Postharvest Biology and Technology*, 191, 111963. <https://doi.org/10.1016/j.postharvbio.2022.111963>
19. Li, Z., Liu, F., Qin, L., & Zhao, Y. (2022). Applications of deep learning in agricultural product quality assessment: A comprehensive review. *Food Control*, 139, 109071. <https://doi.org/10.1016/j.foodcont.2022.109071>
20. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Journal of Intelligent & Robotic Systems*, 91(2), 269–297. <https://doi.org/10.1007/s10846-017-0714-4>
21. Liu, J., Wang, X., Wang, L., Sun, Z., & Zhang, Y. (2021). Hyperspectral imaging for fruit and vegetable quality and safety evaluation: A review. *Comprehensive Reviews in Food Science and Food Safety*, 20(2), 1857–1886. <https://doi.org/10.1111/1541-4337.12703>
22. Lundberg, S. M., & Lee, S.-I. (2020). A unified approach to interpreting model predictions. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
23. Majeed, Y., Lee, W. S., Nam, Y., Choi, S., & Song, K. B. (2020). Deep learning for image-based fruit quality estimation: A review. *Frontiers in Plant Science*, 11, 598604. <https://doi.org/10.3389/fpls.2020.598604>
24. Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Christoph Molnar.
25. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
26. Singh, A., Kaur, H., & Kumar, P. (2024). Explainable AI for smart agriculture: A review and future directions. *Artificial Intelligence in Agriculture*, 12, 100187. <https://doi.org/10.1016/j.aiia.2023.100187>

27. Sriram, R. (2019). *Fruits are fresh and rotten for classification*. Kaggle. <https://www.kaggle.com/datasets/sriramr/fruits-fresh-and-rotten-for-classification>
28. Uçar, F., & Korkmaz, D. (2020). COVIDiagnosis-Net: Deep Bayes-Squeeze Net-based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. *Medical Hypotheses*, 140, 109761. <https://doi.org/10.1016/j.mehy.2020.109761>
29. Wang, J., He, D., & Tong, J. (2022). Apple bruise detection using deep learning: A review. *Journal of Food Process Engineering*, 45(2), e13922. <https://doi.org/10.1111/jfpe.13922>
30. Xie, C., Shao, Y., Li, X., He, Y., & Tu, K. (2023). Recent advances in deep learning for fruit quality inspection: A review. *Trends in Food Science & Technology*, 133, 281–295. <https://doi.org/10.1016/j.tifs.2023.03.011>
31. Yadav, S. K., Kumar, A., Ghosal, S., Joshi, G. P., & Vig, L. (2023). Explainable artificial intelligence (XAI): An emerging paradigm for building trust and transparency in agriculture and food supply chains. *Trends in Food Science & Technology*, 163, 254–270. <https://doi.org/10.1016/j.tifs.2023.04.010>
32. Zhang, H., Qin, J., Wu, W., & Liu, W. (2021). CNN-based detection of fruit defects using hyperspectral imaging. *Computers and Electronics in Agriculture*, 185, 106135. <https://doi.org/10.1016/j.compag.2021.106135>
33. Zhang, Y., Wang, J., Li, B., & Li, Y. (2023). Ensemble learning approaches for crop and food quality prediction: A review. *Artificial Intelligence in Agriculture*, 9, 100130. <https://doi.org/10.1016/j.aiia.2022.100130>
34. Zhou, J., Wu, J., Li, Y., & Li, J. (2021). Machine vision for food quality and safety evaluation: A comprehensive review. *Trends in Food Science & Technology*, 112, 1–15. <https://doi.org/10.1016/j.tifs.2021.02.027>
35. Zhou, Z., Sun, Z., Liu, J., & Zhang, Y. (2023). Nondestructive detection of internal quality of fruits and vegetables using hyperspectral/multispectral imaging: A review. *Critical Reviews in Food Science and Nutrition*, 63(11), 1908–1932. <https://doi.org/10.1080/10408398.2021.1987069>