Non-Destructive Prediction of Fruit Ripeness and Firmness Using Hyperspectral Imaging and Lightweight Machine Learning Models

Phongsakon Mark Konrad[a], Casper Kunstmann-Olsen[b], Jacek Fiutowski[b], Serkan Ayvaz[a,*]

[a] *The Maersk Mc-Kinney Moller Institute, SDU Centre for Industrial Software, University of Southern Denmark, Sønderborg, 6400, Denmark*
[b] *Mads Clausen Institute, SDU NanoSyd, University of Southern Denmark, Sønderborg, 6400, Denmark*

Abstract

Post-harvest fruit quality assessment is essential for reducing food waste, yet reliable non-destructive methods typically depend on expensive hyperspectral cameras and computationally intensive deep learning models. These systems typically require GPU resources, large-scale training data, and domain expertise, limiting their feasibility for many real-world agricultural settings. This study systematically evaluates 20 classical machine learning algorithms on hyperspectral imaging data for simultaneous ripeness classification and firmness prediction across five fruit species, using cross-validated experimental design with Bayesian hyperparameter optimization. Data preprocessing strategy, particularly class balancing and spectral transformations, contributes as much to prediction accuracy as algorithm choice. Our results show that tree-based machine learning models can outperform state-of-the-art deep

*Corresponding author
Email address: seay@mmmi.sdu.dk (Serkan Ayvaz)

learning models reported in Fruit-HSNet. Moreover, the findings indicate that only three visible-range wavelengths are needed to recover over 94% of full-spectrum accuracy, demonstrating that low-cost multispectral sensors combined with lightweight machine learning models can serve as practical alternatives to expensive hyperspectral cameras and complex deep learning approaches for practical fruit quality sorting.

Keywords: Hyperspectral imaging, Machine learning, Multi-attribute Fruit quality assessment, Non-destructive testing, Precision agriculture

1. Introduction

Commercial fruit quality assessment relies on destructive sampling of small fractions of harvests or subjective visual inspection (Lu et al., 2020). A typical processing facility handles tens of thousands of items daily but can destructively test perhaps 100 samples. The remainder is graded by surface appearance, a weak proxy for internal quality that leads to misclassification, post-harvest losses, and inefficient supply-chain decisions.

Hyperspectral imaging (HSI) addresses this gap through non-destructive, contact-free measurement of wavelength-dependent reflectance across the visible and near-infrared spectrum. Spectral signatures encode biochemical and structural changes associated with ripeness, firmness, and spoilage (Yang et al., 2025). Laboratory studies report accuracy exceeding 84% for strawberry maturity assessment (Su et al., 2021), yet commercial adoption remains limited.

Deep learning architectures achieve strong laboratory performance: 3D-CNNs reach 84% accuracy for strawberry classification (Su et al., 2021),

transformer models reach 99.4% for contamination detection (Guo et al., 2024), and real-time systems achieve 98.6% accuracy (Gao et al., 2020). However, these approaches require GPU acceleration, large training datasets, and specialist expertise that most agricultural operations cannot justify. Classical machine learning offers a different set of tradeoffs. These methods train efficiently on limited data, run on consumer-grade CPU hardware, and produce interpretable feature importance scores amenable to validation against plant physiology. Whether classical methods can achieve competitive accuracy on hyperspectral fruit quality benchmarks remains uncharacterized.

Systematic benchmarking is lacking. Published comparisons typically fix a single preprocessing pipeline without ablation, leaving open the question of how much accuracy variation stems from data engineering rather than algorithm selection.

This study benchmarks 20 classical and gradient-boosted machine learning algorithms on paired ripeness and firmness prediction using the DeepHS Fruit dataset (Varga et al., 2021). A six-phase experimental design spans 200 model–configuration pairs, 4,000 Bayesian optimization trials, 10-fold cross-validation, ensemble learning, and explainable AI analysis. The central question is whether preprocessing decisions contribute as much to classification accuracy as algorithm choice, and if so, what mechanisms drive that effect. By evaluating all 20 algorithms under identical, controlled conditions, this benchmark provides a transparent reference point for the field. The main contributions are:

1. Systematic ablation across 200 configurations provides the first evidence that preprocessing shifts accuracy as much as algorithm selection

(25.2 vs. 32.4 pp; bootstrap 95% CI on the difference contains zero).

2. Firmness prediction outperforms ripeness classification across all 20 models (d=1.84), traced to label-structure asymmetry rather than algorithm-specific effects.
3. PCA degrades performance after engineered spectral transforms because variance-maximizing compression discards low-variance features that carry discriminative information.
4. Consensus feature ranking identifies three visible-range wavelengths (448, 540, 640 nm) that recover 94.7% of full-spectrum accuracy; approximate RGB wavelengths (450, 550, 651 nm) perform comparably (96.1%), indicating that band count rather than precise band placement drives the reduced-band performance.

The remainder of this paper is organized as follows. Section 2 reviews prior work on HSI-based fruit quality assessment, preprocessing pipelines, and explainability methods. Section 3 describes the dataset, feature engineering, and six-phase experimental design. Section 4 presents results across all phases, including ablation, cross-validation, and XAI analysis. Section 5 interprets the key findings and discusses limitations. Section 6 summarizes the contributions and outlines directions for future work.

2. Related Work

Three areas of prior work frame the contribution of this study.

2.1. HSI and Machine Learning for Fruit Quality

Hyperspectral systems acquire spectral information across hundreds of narrow wavelength bands, spanning visible (400–700 nm) and near-infrared

(700–2500 nm) regions. This spectral data enables detection of biochemical changes during fruit maturation: chlorophyll degradation, carotenoid accumulation, and cellular structure modifications (Lu et al., 2020). Spectral-biochemical relationships have been established for quality attributes: chlorophyll absorption peaks around 680 nm indicate ripeness state, while near-infrared signatures (850–950 nm) correlate with cellular structure and firmness (Khodabakhshian and Emadi, 2017; Feng et al., 2023). Translating these laboratory findings into deployable systems remains difficult.

Traditional machine learning approaches, including Partial Least Squares Regression (PLSR), Support Vector Machines (SVM), and tree-based ensembles, achieved 70–85% accuracy with advantages in interpretability and computational efficiency (Lu et al., 2017; Khodabakhshian and Emadi, 2017). Reviews of the field identify PLSR and SVM as the most commonly applied methods (Wieme et al., 2022), and targeted wavelength selection can maintain classification accuracy with as few as six bands (Nagasubramanian et al., 2018). These practical insights have received limited attention as the field shifted toward deep learning, and systematic evaluation of traditional methods under standardized conditions remains sparse.

Despite these accuracy gains, deep learning models require GPU acceleration, specialized expertise, and extensive training datasets unavailable for diverse fruit varieties or regional conditions. Computational requirements exceed what most agricultural facilities can justify. Hybrid approaches combining traditional feature engineering with deep learning (Liu et al., 2024; Olisah et al., 2024) often compound complexity without addressing deployment barriers.

2.2. Multi-Attribute Prediction and Explainability

Ripeness and firmness are biologically coupled through ethylene production, cell wall breakdown, and chlorophyll degradation (Feng et al., 2023; Varga et al., 2021), yet most research treats them as independent prediction problems. Feng et al. (2023) found advantages for joint prediction over independent models in loquat quality assessment, and multi-modal approaches combining hyperspectral and RGB data show 10–15% accuracy improvements (Garillos-Manliguez and Chiang, 2021). Yet single-task evaluation persists despite the fact that isolated quality attributes rarely drive commercial sorting decisions. This study adopts a paired single-task evaluation framework assessing ripeness and firmness jointly via aggregated metrics, while training separate models per task.

SHAP and LIME analyses identify biologically meaningful wavelengths: 680 nm for chlorophyll absorption, 760 nm for red-edge transitions, 970 nm for water content (Ahmed et al., 2024b). These findings align with plant physiology knowledge. If only 6–15 wavelengths contain most discriminative information (Nagasubramanian et al., 2018; Ahmed et al., 2024a), full-spectrum systems may be overengineered and targeted multispectral sensors could achieve comparable performance at reduced cost.

2.3. Benchmarking Gaps and Deployment Challenges

Most studies evaluate 2–5 algorithms under carefully selected conditions (Yang et al., 2025). Reported accuracies range from 73% to 99% for similar classification tasks (Vignati et al., 2023). The DeepHS Fruit dataset (Varga et al., 2021) provides standardized data across five fruit varieties (avocado,

kiwi, mango, kaki/persimmon, papaya) with ripeness and firmness annotations. Ben Jmaa et al. (2025) introduced Fruit-HSNet, reporting 70.73% overall accuracy on this benchmark. Systematic evaluation of traditional machine learning methods on this benchmark remains limited, with most comparative studies focusing on deep learning architectures.

Hyperspectral systems cost $10,000–100,000 compared to $500–5,000 for RGB alternatives, require specialized operators, and experience 15–30% accuracy degradation under variable field conditions (Min et al., 2023). Processing speed requirements for sorting applications (under 100 ms per sample) eliminate most deep learning approaches. RGB-reconstructed hyperspectral images (Ahmed et al., 2024a) and edge computing solutions (Lanke and Chandak, 2025) address symptoms rather than causes.

No existing study systematically benchmarks a broad set of traditional ML algorithms on multi-attribute fruit quality prediction with controlled preprocessing ablation.

3. Materials and Methods

3.1. Dataset Description

We used the DeepHS Fruit dataset (version 2), a publicly available hyperspectral imaging dataset for fruit quality assessment research (Varga et al., 2021). The dataset is hosted on GitHub[1] and provides hyperspectral data with quality annotations for benchmarking. Version 2 corrects a potential test-set contamination issue in the original split by ensuring that balanced

[1]https://github.com/cogsys-tuebingen/deephs_fruit

training variants do not alter the benchmark test partition; train, validation, and test sets are validated for zero sample-level overlap.

We use the dataset-provided VIS benchmark split as our fixed evaluation setting (unbalanced train/test: train n=381, test n=138; total n=519). Fruit-type distribution in this split is avocado (n=170, 32.8%), kiwi (n=162, 31.2%), mango (n=68, 13.1%), kaki/persimmon (n=68, 13.1%), and papaya (n=51, 9.8%). The VIS subset spans 47 acquisition days, providing temporal variation in fruit maturation states. For the stratified resplit condition, we keep the benchmark test set fixed but replace the training set with a fruit-balanced training set (n=414; avocado n=130, kiwi n=130, mango n=56, kaki/persimmon n=55, papaya n=43), improving balance without breaking comparability.

The full DeepHS Fruit dataset includes multiple hyperspectral imaging systems (Specim FX10, INNO-SPEC Redeye 1.7, and Corning microHSI 410 Vis-NIR). To maintain spectral consistency and comparability with existing benchmarks, this study restricts evaluation to the VIS camera (Specim FX10, 398–1004 nm), and all reported results reflect this sensor and its wavelength range. The FX10 covers only the short-NIR region; true NIR water absorption bands (1200–1450 nm) that are commonly used in firmness assessment are absent from these data.

Each sample includes quality annotations from destructive measurements. Ripeness classification follows a three-class scheme based on visual assessment and destructive sampling (Varga et al., 2021); inter-annotator agreement was not reported in the original dataset publication. Classes are: unripe (n=116, 22.4%), perfect (n=273, 52.6%), and overripe (n=130, 25.0%). Per-

fect ripeness samples are overrepresented at 52.6%, while unripe and overripe samples are more evenly distributed.

Firmness measurements (grams-force from penetrometer testing) are categorized into three classes derived from cross-fruit distribution analysis: soft (0–1000 gf, n=220, 42.4%), medium (1001–2500 gf, n=186, 35.8%), and firm (2501+ gf, n=99, 19.1%). The distribution skews toward softer fruits (42.4% soft vs 19.1% firm). Samples with missing firmness measurements (n=14, 2.7%; train n=13, test n=1) are retained and encoded as a separate "Unknown" label during training to avoid discarding data; they are excluded from all quality-metric computations reported in the tables.

Sample distribution shows fruit-type imbalances (avocado n=170 vs papaya n=51) and ripeness-class imbalances (perfect 52.6% vs unripe 22.4%, Fig. 1a). PCA analysis indicates that 95% cumulative explained variance is achieved with approximately 18–20 principal components from the 1,120-dimensional feature space (Fig. 1b). Projection onto the first two principal components (Fig. 1c) reveals that cluster structure is driven primarily by fruit type rather than quality state. This species-dominated variance structure motivates per-task rather than joint classification approaches.

3.2. Data Preprocessing and Feature Engineering

3.2.1. Spectral Feature Extraction

The DeepHS Fruit dataset provides hyperspectral data from three imaging systems (Specim FX10, INNO-SPEC Redeye 1.7, Corning microHSI 410 Vis-NIR). We use only the VIS camera (Specim FX10; 224 bands, 398–1004 nm) for spectral consistency. We consume the dataset-provided extracted spectral feature tables (derived from background-masked hyperspec-

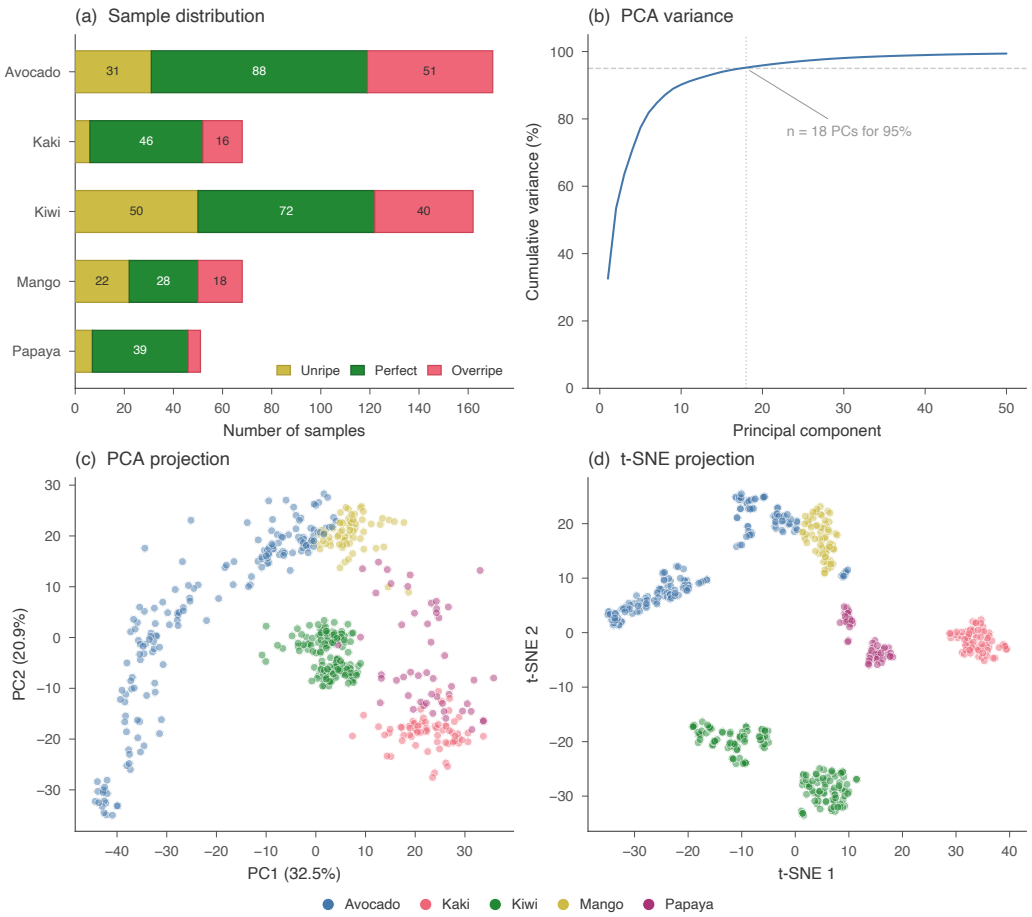

Figure 1: Dataset overview. **(a)** Sample distribution across five fruit types and three ripeness classes. **(b)** PCA cumulative variance: 18 components capture 95% of variance in the 1,120-dimensional feature space. **(c)** PCA projection and **(d)** t-SNE projection colored by fruit type show species-driven cluster structure rather than quality-state separation.

tral cubes to isolate fruit pixels) and do not re-run pixel-level segmentation within the benchmarking pipeline.

An average spectrum is calculated for each sample by averaging spectral values across all foreground (fruit) pixels for each wavelength band:

$$\bar{S}(\lambda) = \frac{1}{|\mathcal{F}|} \sum_{(i,j)\in\mathcal{F}} S_{i,j}(\lambda) \tag{1}$$

where $S_{i,j}(\lambda)$ represents the spectral value at wavelength λ for pixel (i, j), and $\mathcal{F}$ denotes the set of foreground pixels. Each hyperspectral cube reduces to a 224-dimensional spectral signature per sample.

3.2.2. Derived Spectral Features

The combined feature set concatenates five spectral representations from the extracted feature tables, totaling 1,120 dimensions (5 representations $\times$ 224 spectral bands).

The average spectrum $\bar{S}(\lambda)$ captures absolute spectral reflectance across the VIS–NIR range. The first derivative highlights regions of rapid spectral change, emphasizing transitions related to chlorophyll degradation, carotenoid accumulation, and other biochemical changes:

$$D_1(\lambda) = \frac{d\bar{S}(\lambda)}{d\lambda}. \tag{2}$$

Continuum removal normalizes the spectrum by fitting a convex hull and dividing the original spectrum by this hull, highlighting absorption bands such as chlorophyll absorption around 680 nm:

$$CR(\lambda) = \frac{\bar{S}(\lambda)}{C(\lambda)}, \tag{3}$$

where $C(\lambda)$ is the continuum value at wavelength λ.

The standard normal variate (SNV) standardizes each mean spectrum:

$$SNV(\lambda) = \frac{\bar{S}(\lambda) - \mu}{\sigma}, \tag{4}$$

where μ and σ are the mean and standard deviation of $\bar{S}(\lambda)$. Applied to per-sample mean spectra (averaged across pixels), SNV corrects for baseline offset and gain differences between samples; it does not correct within-pixel multiplicative scatter in the sense used for raw reflectance cubes.

Finally, the first derivative of the continuum-removed spectrum combines absorption-emphasizing and edge-enhancing properties:

$$D_1^{CR}(\lambda) = \frac{dCR(\lambda)}{d\lambda}. \tag{5}$$

These five representations capture complementary aspects of spectral variation.

Raw spectral reflectance signatures for the five fruit types exhibit characteristic pigment-related variations across ripeness states, particularly in chlorophyll absorption regions (∼680 nm) and red-edge transitions (∼700–750 nm). Near-infrared features (700–1000 nm) show structure-related variations correlated with firmness. These signatures arise from changes in cellular architecture and water content during maturation (Fig. 2).

3.2.3. Dataset Balancing and Train/Test Splits

The original DeepHS Fruit splits show class imbalances for certain fruit types, ripeness states, and firmness categories. We distinguish two categories of data balancing:

Training set replacement. The stratified resplit condition replaces the training partition with a fruit-balanced set (n=414 vs the original n=381),

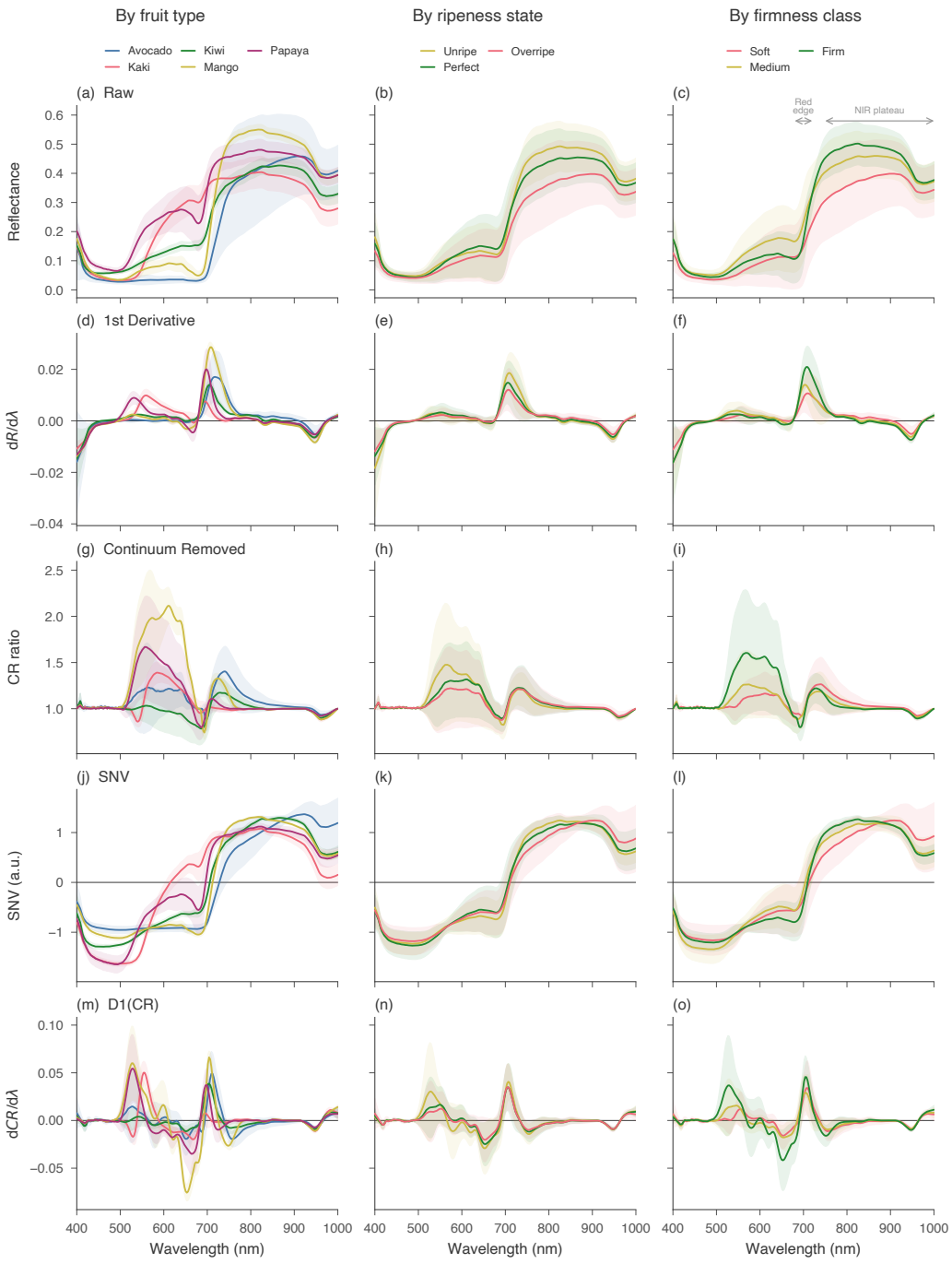

Figure 2: Five spectral preprocessing types (rows) across the 224 bands (400–1000 nm), grouped by fruit type, ripeness state, and firmness class (columns). Shaded bands: mean ± std. Rows: (a–c) raw, (d–f) 1st derivative, (g–i) continuum removed, (j–l) SNV, (m–o) D1(CR). Inter-species variation dominates; class separation is clearest in derivative features near the red edge (∼700 nm).

keeping the benchmark test set fixed. This changes both the sample composition and size (an 8.7% increase in training samples), so its advantage may partly reflect additional data rather than purely improved balance.

Training set augmentation. SMOTE, random oversampling, and random undersampling modify the class distribution within the original training partition without changing which real samples are included. Sampling is applied independently per task (ripeness or firmness); the test partition is kept fixed.

Firmness is discretized into three bins: soft (0–1000 gf), medium (1001–2500 gf), and firm (2501+ gf), derived from cross-fruit firmness distribution analysis. The original DeepHS dataset defines fruit-specific thresholds (e.g., $<900/900$–$1200/>1200\,\mathrm{g/cm^2}$ for avocado; Varga et al. 2021), but universal bins enable direct cross-fruit comparison across the full firmness spectrum.

The original unbalanced split is preserved for comparability with benchmarks from Varga et al. (Varga et al., 2021) and Ben Jmaa et al. (Ben Jmaa et al., 2025).

3.2.4. Additional Preprocessing Strategies

When enabled, PCA reduces dimensionality while retaining 95% of variance, compressing the 1,120-dimensional feature space to approximately 18–20 principal components. PCA is integrated within the scikit-learn Pipeline class: the transformation is fit only on training data and applied to test data, and in Phase 4 cross-validation PCA is fit separately within each fold. The five spectral representations are concatenated before PCA, so it operates on the full combined feature space.

SMOTE (Synthetic Minority Oversampling Technique) generates syn-

thetic samples for minority classes using the imbalanced-learn[2] library (k=5 nearest neighbors), applied independently per task after train/test splitting but before model training. For the firmness task, which includes 14 samples (2.7%) with missing values encoded as a separate class, SMOTE treats this as a fourth firmness category. Random oversampling duplicates minority class samples; random undersampling removes majority class samples.

Phase 2 evaluates these strategies across 200 configurations (5 data balancing strategies × 2 PCA options × 20 models).

3.2.5. Wavelength Subset Experiments (VIS-3 and RGB)

Two three-band configurations are evaluated to assess the necessity of full-spectrum sensing.

VIS-3 (XAI-derived). Three visible-range bands are selected from hyperspectral data at indices $\{18, 52, 89\}$, corresponding to wavelengths $\{448, 540, 640\}$ nm. These bands were identified through consensus feature ranking from ExtraTrees explainability analysis (Section 4), selecting the top joint-ranked visible-range bands (<700 nm) across both ripeness and firmness tasks. We term this the "VIS-3 subset" to emphasize that these bands are selected from hyperspectral cubes, not acquired by an independent RGB sensor.

RGB (approximate). Three bands at indices $\{19, 56, 93\}$, corresponding to wavelengths $\{450, 550, 651\}$ nm, approximate the center wavelengths of standard Bayer-pattern RGB sensors. This configuration tests whether XAI-guided band selection outperforms generic RGB-center wavelengths, thereby isolating the contribution of precise band placement from the effect of spectral

[2]imbalanced-learn: `https://imbalanced-learn.org`

dimensionality reduction.

Both configurations apply the same five spectral transformations, resulting in a 15-dimensional feature vector (3 bands $\times$ 5 methods) compared to the 1,120-dimensional full-spectrum vector. Derivatives and continuum removal of three non-contiguous bands lack the physical meaning they carry over continuous spectra; reduced-band results therefore represent an upper bound on what a true 3-band sensor could achieve, not a realistic sensor comparison. Both pipelines follow identical experimental procedures to the full-spectrum pipeline.

3.3. Experimental Design

The six-phase methodology trains each algorithm separately for ripeness and firmness, with performance reported jointly through aggregated metrics.

3.3.1. Pipeline Phases

Each model is trained separately for ripeness and firmness prediction, with performance evaluated using overall accuracy (OA, Eq. 11). We term this "paired single-task evaluation" to distinguish it from multi-task learning architectures that share representations across tasks. Each task is modeled independently; only the evaluation metric aggregates performance. We use separate models rather than joint multi-output classification for four reasons: (1) many evaluated algorithms do not natively support multi-output classification; (2) separate models allow independent hyperparameter tuning (firmness accuracy exceeds ripeness accuracy by approximately 25 pp, Table 5); (3) agricultural operations may prioritize one attribute depending on market demands; and (4) existing literature predominantly uses single-task

evaluation. This approach does not exploit shared representations that joint models might capture; future work should compare separate-task versus joint multi-task architectures.

Phase 1 establishes baseline performance for 20 algorithms on the original imbalanced dataset using default hyperparameters, evaluated using accuracy, F1-macro, and F1-weighted metrics for each task. Phase 2 evaluates preprocessing strategies using a full-factorial design across data balancing strategy (five levels: original imbalanced, stratified resplit, SMOTE, random oversampling, random undersampling), PCA dimensionality reduction (enabled at 95% variance or disabled), and model selection (20 algorithms), generating 200 configurations. The best Phase 2 preprocessing setting per modality is selected for Phase 3 based on overall accuracy. This selection uses test-set performance, introducing indirect information leakage: the chosen preprocessing was optimized, in part, on the same held-out set used for final evaluation. Phase 4 cross-validation provides a partial mitigation, but the preprocessing choice itself was not cross-validated.

Phase 3 applies Bayesian optimization via Optuna[3] (Akiba et al., 2019) with the best Phase 2 preprocessing configuration. For each model, 100 trials using Tree-structured Parzen Estimator (TPE) sampling (Bergstra et al., 2011) maximize $\overline{\text{F1}}_{\text{macro}}$ (Eq. 12). Search spaces are model-specific (Table B.1): tree-based models optimize depth, minimum samples per leaf, and ensemble size; linear models optimize regularization; neural networks optimize learning rate, hidden layer sizes, and activations. Optimization uses

[3]Optuna: `https://optuna.org`

stratified 70/30 train/validation splits, with final models retrained on the full training set. Total: 2,000 trials per modality (20 models × 100 trials), 4,000 trials across both modalities.

Phase 4 validates model performance through 10-fold stratified cross-validation using optimal hyperparameters from Phase 3 and best preprocessing from Phase 2. Results are reported as mean ± standard deviation across folds, with 95% confidence intervals and coefficient of variation. Phases 1–3 provide exploratory results on a single fixed split; Phase 4 cross-validation is the authoritative basis for algorithm ranking and all primary conclusions.

Phase 5 combines top models from Phase 4 using soft voting, hard voting, stacking, and blending. Phase 6 performs dedicated explainability analysis for the best-performing single model using SHAP[4] (Lundberg and Lee, 2017), LIME[5] (Ribeiro et al., 2016), and permutation importance (Breiman, 2001), producing wavelength-level interpretations for both tasks. The full pipeline architecture is illustrated in Fig. 3.

3.3.2. Benchmark Reference

We reference Fruit-HSNet's reported 70.73% overall accuracy (Ben Jmaa et al., 2025) as a contextual data point, not a controlled comparison. Fruit-HSNet processes full 64×64 pixel hyperspectral cubes through a dual-branch network that extracts Fourier-domain spectral features alongside central-pixel spatial signatures, trained with GPU acceleration, focal loss, and spatial data augmentation including random cropping, flipping, rotation, and noise

[4]SHAP: `https://github.com/shap/shap`

[5]LIME: `https://github.com/marcotcr/lime`

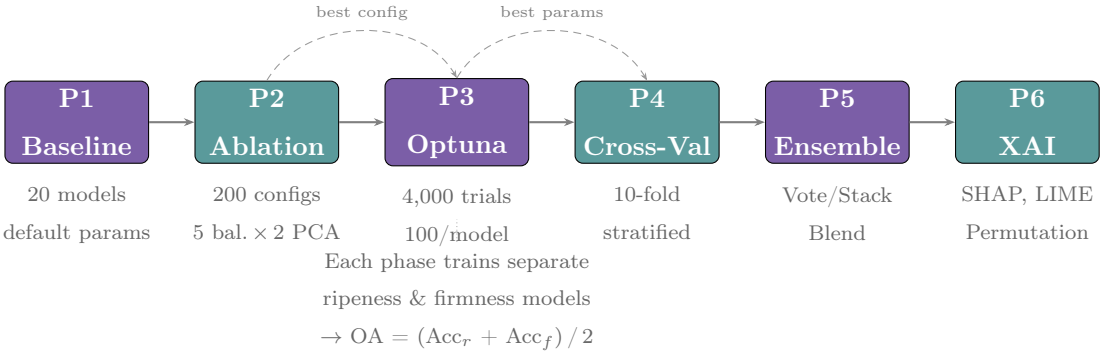

Figure 3: Six-phase benchmark pipeline. Each phase trains independent ripeness and firmness models, combined via overall accuracy (OA). Dashed arrows indicate configuration propagation between phases.

injection. Our approach discards spatial information entirely, representing each sample by its per-pixel-averaged mean spectrum (a 224-dimensional vector retaining no spatial structure) processed on a consumer CPU in under one second. The following protocol differences preclude direct comparison: (1) Fruit-HSNet used a different train/test partition, (2) label preprocessing and class definitions may differ, and (3) Fruit-HSNet is a deep learning architecture while this study evaluates the scikit-learn ecosystem. Our Phase 3 result (ExtraTrees, 75.0%) carries a Wilson 95% confidence interval of [67.6%, 81.8%], which contains Fruit-HSNet's reported value, and no statistically meaningful difference can be claimed.

Phase 4 cross-validation provides statistical estimates through 10-fold stratified sampling (mean, standard deviation, coefficient of variation, 95% confidence intervals). We apply the Friedman test for omnibus comparison of all 20 classifiers on per-fold F1 scores, followed by Nemenyi post-hoc tests with family-wise error control (Table 2). Cohen's d effect sizes quantify task

asymmetry (Table 3). Phase 4 cross-validated rankings are the authoritative ranking for algorithm comparison; Phase 3 single-split results are retained only for protocol-matched reference against Fruit-HSNet. Reproducibility details are provided in Table B.3.

3.4. Machine Learning Models

We evaluate 20 classical and gradient-boosted machine learning algorithms from seven algorithmic families, including PLS-DA (Partial Least Squares Discriminant Analysis), the standard chemometric baseline for spectral classification. PLS-DA is implemented as a custom scikit-learn wrapper around `PLSRegression` with dummy-coded class labels and argmax prediction; `n_components` is clamped at fit time to $\min(n_{\text{samples}}, n_{\text{features}}, n_{\text{classes}})$. Note that the sklearn `MLPClassifier` used here is a shallow multi-layer perceptron; it shares an API with other scikit-learn classifiers but is not a deep learning model. Models are implemented using scikit-learn[6] (Pedregosa et al., 2011), XGBoost[7] (Chen and Guestrin, 2016), and LightGBM[8] (Ke et al., 2017). Each model is integrated into a pipeline with feature scaling (StandardScaler) and optional PCA. Table B.1 provides a structured overview of all models, including abbreviations, family membership, hyperparameter counts, Optuna search spaces, and native feature importance support.

All models use fixed random state (seed=42), early stopping where applicable, and class balancing for imbalanced datasets. A complete overview of all models with Optuna search spaces is provided in Table B.1 (Appendix);

[6]scikit-learn: `https://scikit-learn.org`

[7]XGBoost: `https://xgboost.ai`

[8]LightGBM: `https://lightgbm.readthedocs.io`

best hyperparameters from Bayesian optimization are listed in Table B.2.

3.5. Evaluation Metrics

We report per-task accuracies for ripeness (r) and firmness (f), overall accuracy (OA) defined as the arithmetic mean of task accuracies, and F1-scores (both macro-averaged and weighted). Per-class metrics for class c:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \tag{6}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \tag{7}$$

$$\text{F1}_c = \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{8}$$

Task-level aggregates for task $t \in \{r, f\}$ with class set C_t:

$$\text{F1}^{(t)}_{\text{macro}} = \frac{1}{|C_t|} \sum_{c \in C_t} \text{F1}_c, \tag{9}$$

$$\text{F1}^{(t)}_{\text{wt}} = \sum_{c \in C_t} \frac{n_c}{N_t}\,\text{F1}_c. \tag{10}$$

Overall metrics average across both tasks:

$$\text{OA} = \frac{\text{Acc}^{(r)} + \text{Acc}^{(f)}}{2}, \tag{11}$$

$$\overline{\text{F1}}_{\text{macro}} = \frac{\text{F1}^{(r)}_{\text{macro}} + \text{F1}^{(f)}_{\text{macro}}}{2}. \tag{12}$$

Equal weighting treats both quality attributes as equally relevant for holistic assessment; deployment scenarios prioritizing one attribute can reweight accordingly. OA is the primary benchmark comparison metric; $\overline{\text{F1}}_{\text{macro}}$ is the optimization objective for Phase 3 and Phase 4 cross-validation. Cross-validation reports mean ± standard deviation, coefficient of variation, and 95% confidence intervals. Efficiency metrics include training time (s), inference time (s), and model size (MB).

3.6. Computational Environment

All experiments were conducted on consumer-grade hardware, confirming feasibility on consumer hardware without specialized infrastructure (Table B.3). Hardware configuration consists of MacBook Air M4 (2024) with Apple Silicon M4 chip, 24 GB unified memory, SSD storage, and macOS (arm64). Software stack includes Python 3.13, scikit-learn 1.7.1, XGBoost 3.0.4, LightGBM 4.6.0, Optuna 4.5.0, and imbalanced-learn 0.14.0.

CO_2 emissions were negligible across all models (<0.1 g per model on CPU-only consumer hardware; estimated via CodeCarbon[9] methodology (Schmidt et al., 2021)).

4. Results

Evaluation spans 20 algorithms across three modalities (full-spectrum: 1,120 features; VIS-3 subset: 15 features; RGB approximate: 15 features) and six experimental phases, including 10-fold cross-validation and 4,000 hyperparameter-optimization trials. A Friedman test on per-fold overall F1 scores confirmed significant differences among classifiers ($\chi^2(19) = 144.8$, $p < 10^{-20}$); the Nemenyi post-hoc critical difference diagram (Fig. A.1 in Appendix, Table 2) identifies 46 significantly different pairs.

4.1. Pipeline Performance

Table 1 presents integrated results across all pipeline phases for the 20 evaluated models, sorted by Phase 4 cross-validated overall accuracy. XGBoost achieved the highest cross-validated OA ($82.3 \pm 5.4\%$), with HistGra-

[9]CodeCarbon: `https://codecarbon.io`

dientBoosting, LGBM, and ExtraTrees within 0.1 percentage points (82.2% each). In total, 11 of 20 models exceeded 75% OA in 10-fold CV. ExtraTrees, which ranked first on the single benchmark split (Phase 3, 75.0%), maintained strong cross-validated performance. Phase 4 cross-validated rankings are the authoritative comparison. Subsequent per-class and XAI analyses use ExtraTrees because it provides native feature importance scores and achieved the highest single-split accuracy on the benchmark partition used for FruitHSNet comparison. Training time was 0.2 seconds on consumer hardware.

Preprocessing strategy substantially influenced performance for top models. Stratified resplit balancing improved ExtraTrees from 71.7% (Phase 1 baseline) to 75.0% (Phase 2), a 3.3 percentage point gain from data balancing alone. Phase 3 Bayesian optimization (100 trials per model) did not further improve ExtraTrees. For competitive models, preprocessing rather than hyperparameter tuning drove the performance gain (Fig. 4a). All 20 models trained in under 19 seconds on the consumer-grade hardware (MacBook Air M4). Inference latency (Phase 4) is well below the 100 ms sorting-line threshold for all top models: ExtraTrees 2.1 ms/sample, RandomForest 2.9 ms/sample, LGBM 6.8 ms/sample, XGBoost 34.1 ms/sample (138 test samples). The most efficient model (ExtraTrees) is 90× faster than GradientBoosting while achieving 11.2 percentage points higher accuracy (Fig. 4b).

The Pareto frontier in model size vs. accuracy space (Fig. 4b, star markers) identifies six non-dominated models, namely ExtraTrees (75.0%, 2.8 MB), RandomForest (73.6%, 1.4 MB), XGBoost (65.6%, 0.7 MB), GradientBoosting (63.8%, 0.4 MB), AdaBoost (62.0%, 0.1 MB), and DecisionTree (58.7%,

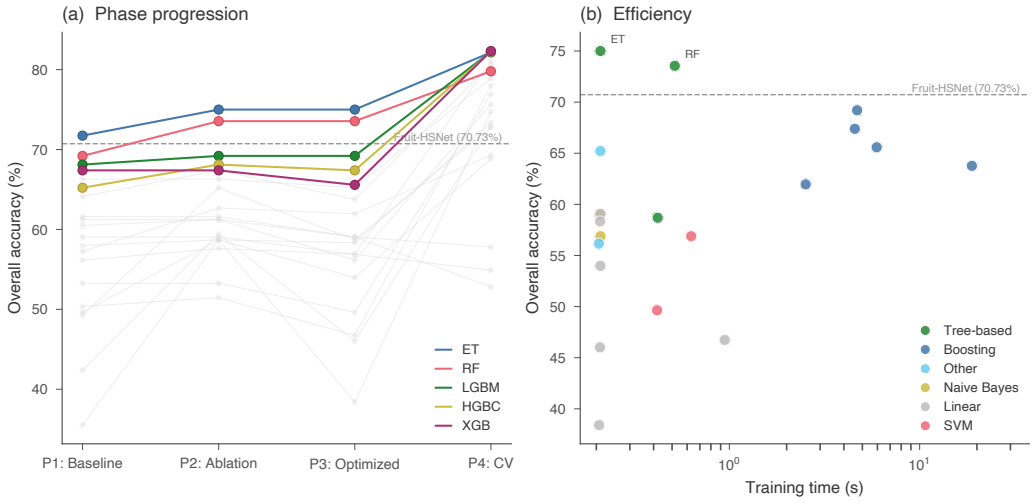

Figure 4: Pipeline performance. (a) Phase progression for all 20 models: connected dots trace each model's overall accuracy from baseline (P1) through ablation (P2, best config), optimization (P3), and 10-fold cross-validation (P4). Top-5 models highlighted; remaining in gray. The largest gains occur between P1 and P2, confirming that preprocessing contributes more than hyperparameter tuning for competitive models. (b) Training-time efficiency (log scale) vs overall accuracy (Phase 3), colored by algorithmic family. Marker size reflects model storage size. Star markers denote Pareto-optimal models in model size vs accuracy space (6 models: ExtraTrees, RandomForest, XGBoost, GradientBoosting, AdaBoost, DecisionTree), representing the best size-accuracy tradeoff frontier. Tree-based methods achieve the highest accuracy with the shortest training times.

0.07 MB). ExtraTrees achieves the highest accuracy at a storage footprint suitable for embedded sorting hardware, with inference latency of 2.1 ms per sample. The dominance of preprocessing over optimization across competitive models raises the question of which preprocessing decisions matter most and why.

Table 1: Model performance and efficiency across pipeline phases (full-spectrum, 1,120 features). Models sorted by Phase 4 cross-validated overall accuracy.[11] P4 shows 10-fold stratified cross-validation accuracy (mean ± std). P3 Wilson 95% CI (binomial, n=138) shown in brackets. **Bold** = best in column; underline = second-best. Arrows indicate preferred direction.

Rk	Model	Family	P1↑ OA	P2↑ OA	P2 Config	P3↑ OA	P3 CI [95%]	P4 CV (%)↑	Time↓ (s)	Size↓ (MB)
1	XGBoost	Boosting	67.4	67.4	SMOTE / No PCA	65.6	[57.0, 72.7]	**82.3** ± 5.4	5.9	0.72
2	HistGradientBoosting	Boosting	65.2	68.1	SMOTE / No PCA	67.4	[59.2, 74.6]	82.2 ± 5.4	4.6	5.05
3	LGBM	Boosting	68.1	69.2	Strat. Resplit / No PCA	69.2	[61.4, 76.6]	82.2 ± 4.9	4.7	1.64
4	ExtraTrees	Tree-based	**71.7**	**75.0**	Strat. Resplit / No PCA	**75.0**	[67.6, 81.8]	82.2 ± 4.7	**0.2**	2.77
5	SVC_RBF	SVM	56.2	57.6	SMOTE / No PCA	56.9	[48.2, 64.5]	81.1 ± 4.0	0.6	5.16
6	GradientBoosting	Boosting	64.1	67.4	Oversample / No PCA	63.8	[55.5, 71.3]	80.6 ± 5.8	18.8	0.41
7	RandomForest	Tree-based	69.2	73.6	Strat. Resplit / No PCA	73.6	[66.0, 80.5]	79.8 ± 4.8	0.5	1.44
8	KNeighbors	Other	66.3	66.3	Strat. Resplit / PCA	65.2	[57.0, 72.7]	78.9 ± 4.4	**0.2**	7.14
9	LogisticRegression	Linear	50.4	51.4	Oversample / No PCA	46.7	[38.3, 54.7]	78.0 ± 5.2	0.9	0.11
10	SVC_Linear	SVM	53.3	53.3	Original / No PCA	49.6	[41.1, 57.5]	76.9 ± 5.2	0.4	3.64
11	MLPClassifier	Other	61.2	61.2	Original / No PCA	56.2	[48.2, 64.5]	75.6 ± 6.0	**0.2**	5.22
12	LDA	Linear	35.5	58.7	Strat. Resplit / PCA	38.4	[30.7, 46.7]	74.7 ± 4.6	**0.2**	0.23
13	SGDClassifier	Linear	58.0	58.7	Strat. Resplit / PCA	58.3	[49.6, 65.9]	73.4 ± 4.2	**0.2**	0.11
14	DecisionTree	Tree-based	49.3	65.2	SMOTE / No PCA	58.7	[50.4, 66.6]	72.8 ± 5.2	0.4	0.07
15	Ridge	Linear	42.4	59.4	Strat. Resplit / PCA	46.0	[38.3, 54.7]	72.8 ± 3.9	**0.2**	0.11
16	AdaBoost	Boosting	57.2	62.7	Original / PCA	62.0	[54.0, 70.0]	69.3 ± 5.9	2.5	0.12
17	PassiveAggressive	Linear	49.6	58.7	SMOTE / PCA	54.0	[45.3, 61.7]	68.8 ± 5.4	**0.2**	0.11
18	PLSDA	Chemometric	61.6	61.6	Original / No PCA	59.1	[51.1, 67.3]	57.8 ± 4.5	**0.2**	0.42
19	BernoulliNB	Naive Bayes	59.1	59.1	Undersample / No PCA	56.9	[48.2, 64.5]	54.9 ± 5.5	**0.2**	0.17
20	GaussianNB	Naive Bayes	60.5	61.2	SMOTE / PCA	59.1	[51.1, 67.3]	52.8 ± 4.9	**0.2**	0.17

Table 2: Model rankings from 10-fold cross-validation (Friedman/Nemenyi, $\alpha = 0.05$). Rank = mean rank across folds (lower is better). Models connected by a bar in Fig. A.1 are not significantly different.

Rank	Model	Avg. Rank↓	Mean F1↑	Family
1	ET	**3.7**	**0.792 ± 0.077**	Tree-based
2	SVM-R	4.7	0.787 ± 0.077	SVM
3	KNN	5.4	0.771 ± 0.061	Other
4	GBC	5.8	0.760 ± 0.064	Boosting
5	RF	6.2	0.758 ± 0.063	Tree-based
6	LGBM	6.2	0.755 ± 0.049	Boosting
7	HGBC	6.9	0.752 ± 0.048	Boosting
8	SVM-L	7.2	0.747 ± 0.067	SVM
9	LR	7.5	0.750 ± 0.077	Linear
10	XGB	7.6	0.743 ± 0.063	Boosting
11	MLP	7.7	0.725 ± 0.049	Other
12	LDA	12.2	0.671 ± 0.060	Linear
13	DT	13.0	0.659 ± 0.052	Tree-based
14	SGD	14.0	0.639 ± 0.055	Linear
15	Ridge	14.6	0.626 ± 0.041	Linear
16	PA	15.5	0.610 ± 0.057	Linear
17	Ada	15.7	0.609 ± 0.044	Boosting
18	BNB	18.1	0.526 ± 0.065	Naive Bayes
19	PLS-DA	18.7	0.491 ± 0.065	Chemometric
20	GNB	19.3	0.486 ± 0.049	Naive Bayes

Table 3: Summary of statistical tests and descriptive comparisons. Effect sizes follow Cohen's conventions: $|d| < 0.2$ negligible, 0.2–0.5 small, 0.5–0.8 medium, > 0.8 large. Friedman and Nemenyi tests are appropriate because models share the same CV folds. Firmness vs. ripeness comparisons involve the same 20 models trained on shared data; formal significance tests are not used because the observations are not independent. Phase 3 modality comparisons (HSI vs. VIS-3, HSI vs. RGB, VIS-3 vs. RGB) use paired Wilcoxon signed-rank tests on per-model overall accuracy.

Comparison	Summary	d	Effect
Friedman (20 classifiers)	$\chi^2(19) = 144.8$, $p < 10^{-20}$	—	—
Nemenyi pairs (α=0.05)	46 of 190 pairs differ	—	—
Firmness > Ripeness (IQR: 21.4–32.8 pp; exception: LDA)	19/20 models; mean gap 25.0 pp	1.84	Large
HSI vs. RGB	Cohen's d from Phase 4 fold F1	1.17	Large
HSI vs. VIS-3	Phase 3 OA, 20 models; $p = 0.46$	0.32	Small
HSI vs. RGB	Phase 3 OA, 20 models; $p = 0.39$	0.14	Negl.
VIS-3 vs. RGB	Phase 3 OA, 20 models; $p = 0.30$	0.27	Small

4.2. Preprocessing Ablation

Table 4 summarizes the Phase 2 ablation study across 200 configurations (5 balancing strategies × 2 PCA options × 20 models). Stratified resplit achieved the highest mean overall accuracy (59.08% without PCA), while random undersampling performed worst (46.56%).

PCA degraded performance in four of five balancing conditions. Across all 200 configurations, models without PCA outperformed those with PCA by 2.46 percentage points on average (55.62% vs 53.16%, Fig. 5). Only 7 of 20 models utilized PCA in their best Phase 2 configuration, and these were predominantly linear models (Ridge, LDA, SGDClassifier) rather than tree-based methods. The full 200-configuration landscape is visualized in Fig. 5, where the stratified resplit / no PCA column consistently achieves the warmest values.

The consistent PCA degradation is surprising given its routine use in hyperspectral analysis and warrants mechanistic explanation.

Across the 20×10 Phase 2 accuracy matrix, the median accuracy range attributable to preprocessing (varying balancing strategy and PCA within each model) is 25.2 pp, compared with 32.4 pp for algorithm selection (varying the model within each preprocessing configuration). A bootstrap analysis (10,000 resamples) yields a 95% CI of [−5.8, 21.9] pp on the difference between these median ranges, an interval that contains zero. Preprocessing and algorithm choice are thus of comparable magnitude on this benchmark. Ridge exemplifies the upper bound of preprocessing impact: its overall accuracy rose from 42.4% (Phase 1 baseline) to 59.4% (Phase 2 best), a 17.0 pp gain from data engineering alone.

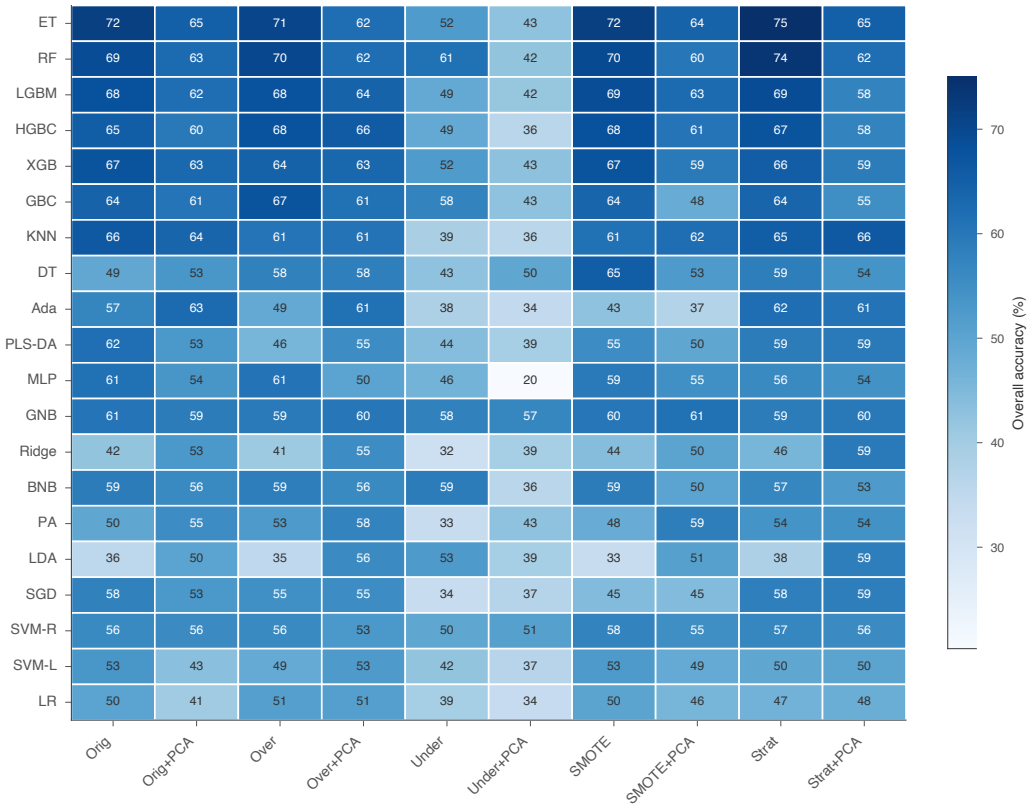

Figure 5: Phase 2 preprocessing ablation across 200 configurations. Overall accuracy (%) for 20 models (rows, sorted by best OA) across 10 preprocessing configurations (5 balancing strategies $\times$ 2 PCA settings). Vertical white lines separate balancing strategies; within each pair, the left column is without PCA and the right is with PCA. The stratified resplit / no PCA column consistently achieves the highest values.

Table 4: Phase 2 ablation: mean ± SD overall accuracy (%)↑ across 20 models by balancing strategy and PCA setting. Δ = no-PCA minus PCA difference. Best strategy in bold. Large SDs (4–11 pp) reflect substantial model-to-model variance within each configuration. †Stratified resplit replaces the training set (n=414 vs n=381); strategies below the rule augment the original training set.

Balancing strategy	No PCA	PCA	Δ (pp)
Stratified resplit†	**59.08±9.26**	**57.41±4.57**	+1.67
SMOTE	57.14±10.61	53.84±7.18	+3.30
Original unbalanced	58.32±9.38	56.29±6.59	+2.03
Random oversampling	57.01±9.85	58.12±4.47	−1.11
Random undersampling	46.56±8.93	40.13±7.47	+6.43
Overall mean	55.62	53.16	+2.46

4.3. Task Asymmetry and Modality Comparison

Table 5 and Fig. 6 reveal two consistent patterns across all 20 models: (1) firmness prediction outperforms ripeness by a wide margin (Fig. 6a), and (2) full-spectrum data outperforms VIS-3 and RGB subsets for most competitive models, but the advantage is small (Fig. 6b).

Firmness accuracy exceeded ripeness accuracy in 19 of 20 models (mean gap 25.0 percentage points; IQR: 21.4–32.8 pp; Cohen's $d = 1.84$, large effect). This asymmetry was most extreme for SVC_RBF (31.9% ripeness vs 81.9% firmness, a 50.0 pp gap) and smallest for SVC_Linear (44.9% vs 54.3%, a 9.4 pp gap). LDA was the only model where ripeness accuracy exceeded firmness, likely reflecting model-specific failure modes. Because the 20 models share a common training set and test set, formal significance tests

for this comparison are not appropriate (Table 3).

Phase 4 fold-level F1 comparisons between HSI and RGB revealed a large full-spectrum advantage (Cohen's $d = 1.17$; Table 3), concentrated among the top-performing models. In Phase 3, VIS-3 and RGB yield comparable performance (mean OA 61.7% vs 60.1%; Cohen's $d = 0.27$, $p = 0.30$), with neither modality significantly outperforming the other (Table 3). Both reduced-band configurations outperformed full-spectrum on average across all 20 models (61.7% and 60.1% vs 59.1%), but this average is driven by weak models (Ridge, LDA, and SVC_Linear) that perform substantially better in the low-dimensional 15-feature space, likely because dimensionality reduction mitigates the multicollinearity that destabilizes their coefficient estimates. At the top of the leaderboard, the full-spectrum advantage persists: ExtraTrees achieves 75.0% (full) vs 72.1% (RGB) vs 71.0% (VIS-3). The persistence of the firmness–ripeness gap across all 20 models and all three modalities suggests a property of the labels rather than of the algorithms.

Table 5: Task asymmetry and modality comparison of Phase 3 models. RA = ripeness accuracy, FA = firmness accuracy, OA = overall accuracy. Full-spectrum has 1,120 features; VIS-3 and RGB each use 15 features. ΔOA = Full OA − max(VIS-3, RGB).

		Full-spectrum			Reduced-band OA↑		
Rk	Model	RA↑	FA↑	OA↑	VIS-3	RGB	ΔOA
1	ExtraTrees	**63.0**	**87.0**	**75.0**	**71.0**	**72.1**	+2.9
2	RandomForest	60.9	86.2	73.6	67.8	71.4	+2.2
3	LGBM	58.0	80.4	69.2	68.1	68.8	+0.4
4	HistGradBoosting	54.3	80.4	67.4	67.4	69.9	−2.5
5	XGBoost	52.2	79.0	65.6	68.5	68.5	−2.9
6	KNeighbors	48.6	81.9	65.2	65.6	66.7	−1.4
7	GradientBoosting	47.1	80.4	63.8	67.4	65.9	−3.6
8	AdaBoost	48.6	75.4	62.0	60.9	44.9	+1.1
9	GaussianNB	44.2	73.9	59.1	54.0	52.5	+5.1
10	PLSDA	39.9	78.3	59.1	51.8	48.6	+7.2
11	DecisionTree	50.7	66.7	58.7	58.3	60.9	−2.2
12	SGDClassifier	42.0	74.6	58.3	60.9	57.2	−2.5
13	BernoulliNB	41.3	72.5	56.9	60.5	53.3	−3.6
14	SVC_RBF	31.9	81.9	56.9	61.2	58.7	−4.3
15	MLPClassifier	39.1	73.2	56.2	51.8	58.0	−1.8
16	PassiveAggr.	42.0	65.9	54.0	58.7	60.1	−6.2
17	SVC_Linear	44.9	54.3	49.6	64.1	57.2	−14.5
18	LogisticRegr.	37.7	55.8	46.7	47.5	56.5	−9.8
19	Ridge	37.0	55.1	46.0	66.7	56.5	−20.7
20	LDA	47.8	29.0	38.4	61.2	54.7	−22.8
Mean		46.6	71.6	59.1	61.7	60.1	-4.0

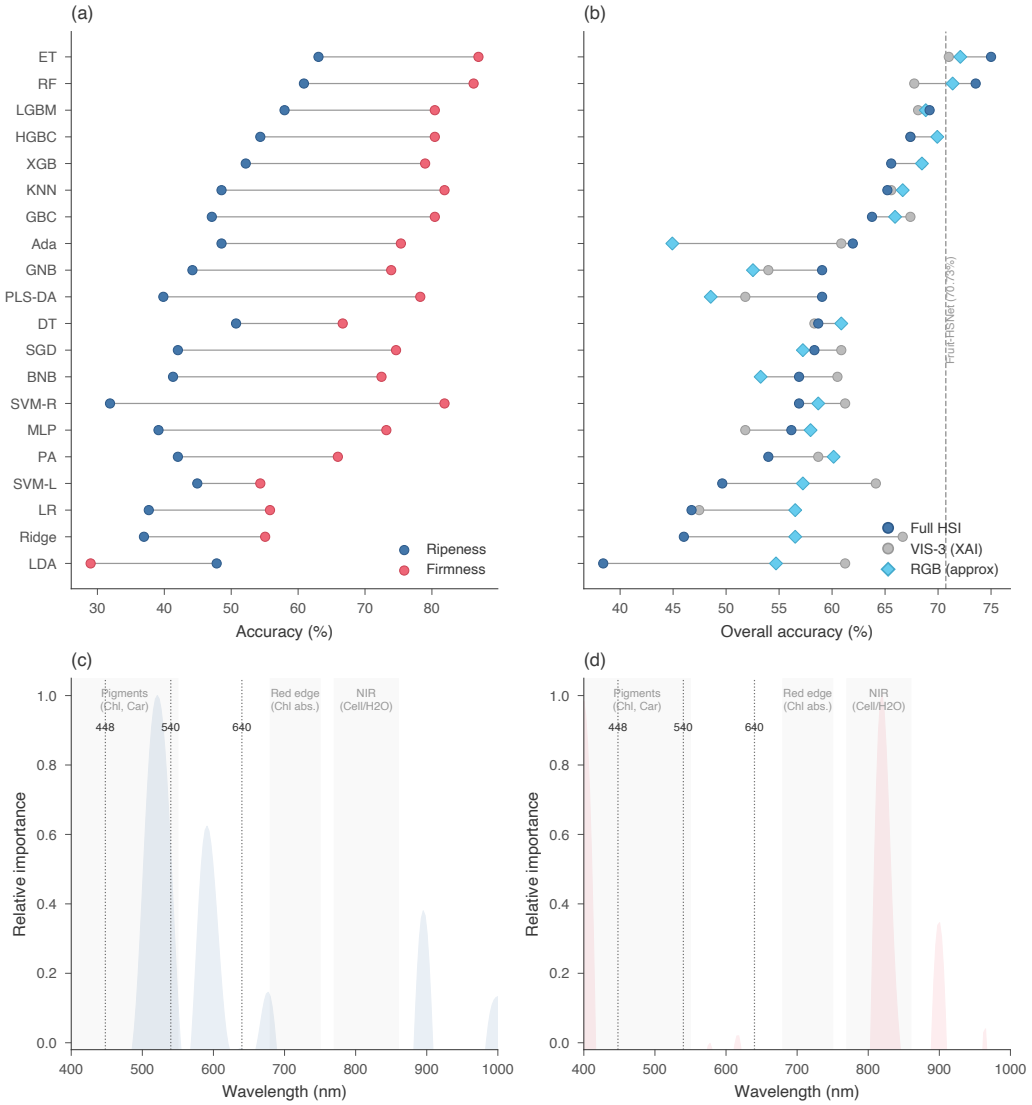

Figure 6: Task asymmetry, modality comparison, and spectral importance for Phase 3 optimized models. **(a)** Task asymmetry: ripeness (pink) vs firmness (green) accuracy per model, revealing a persistent gap across all 20 algorithms. **(b)** Modality comparison: full-spectrum HSI (blue) vs VIS-3 subset (orange) overall accuracy, with the Fruit-HSNet benchmark (70.73%) shown as a dashed line. **(c, d)** Spectral band importance (rolling mean ± std) for ExtraTrees across 400–1000 nm; shaded regions highlight biologically meaningful zones (pigment changes, red-edge chlorophyll absorption, NIR cellular/water signatures). Ripeness importance concentrates in visible wavelengths (c), while firmness importance shifts toward NIR features (d).

4.4. Per-Fruit, Per-Class, and Ensemble Analysis

Table 6 reports per-fruit accuracy for the top five models (Fig. A.3 in Appendix), revealing substantial variation across species. Avocado and mango achieved highest accuracies (up to 75.0% for mango with RandomForest), while papaya consistently underperformed (33.3–44.4%), likely due to its smallest sample size (n=51).

Table 6: Per-fruit ripeness accuracy (%) for the top-five models (Phase 4 single-split holdout, n=138; ranked by Phase 3 overall accuracy). **Bold** = best per fruit; underline = second-best.

Fruit (n)	ExtraTrees	RandomForest	LightGBM	HistGradBoosting	XGBoost
Avocado (170)	**67.9**	65.4	44.4	48.1	38.3
Kaki (68)	**41.7**	**41.7**	33.3	**41.7**	**41.7**
Kiwi (162)	58.3	54.2	62.5	**70.8**	50.0
Mango (68)	66.7	**75.0**	41.7	66.7	50.0
Papaya (51)	33.3	33.3	33.3	33.3	**44.4**

Table 7 provides per-class precision, recall, and F1-score from the single-split holdout evaluation. Ripeness classification showed primary confusion between adjacent states (unripe/perfect and perfect/overripe), as visible in the normalized confusion matrix (Fig. 7a). Stratifying ripeness accuracy by firmness group reveals that extreme-firmness samples (soft: 74.1%, very firm: 73.7%) are classified more accurately than intermediate groups (medium: 54.4%, firm: 56.5%; Table 7, Fig. 7b), suggesting that fruits at firmness extremes present clearer ripeness signatures.

Ten-fold stratified cross-validation (Fig. 8, Fig. A.2 in Appendix) con-

firmed that top models generalize reliably. ExtraTrees achieved $82.2 \pm 4.7\%$ mean OA, with narrow 95% confidence intervals. Firmness accuracy was consistently higher than ripeness across all models, with tighter variance. This pattern reinforces the task asymmetry observed in Phase 3.

Table 7: ExtraTrees Phase 4 single-split holdout analysis (n=138). **Top**: ripeness per-class metrics from the confusion matrix (Fig. 7a). **Bottom**: ripeness accuracy stratified by firmness group (4-bin analysis on the original unbalanced split). Extreme-firmness groups achieve higher ripeness accuracy than intermediate groups. *Unripe precision (1.000) with 47.1% recall means the model rarely predicts unripe; most unripe fruit are misclassified as perfect (41.2%) or overripe (11.8%).

Task	Class	Precision	Recall	F1
Ripeness	Overripe	0.561	**0.696**	0.621
Ripeness	Perfect	0.569	0.638	0.602
Ripeness	Unripe*	**1.000**	0.471	**0.640**
Firmness group		*Ripeness accuracy*		
	Soft (n=27)	0.741		
	Medium (n=68)	0.544		
	Firm (n=23)	0.565		
	Very firm (n=19)	**0.737**		

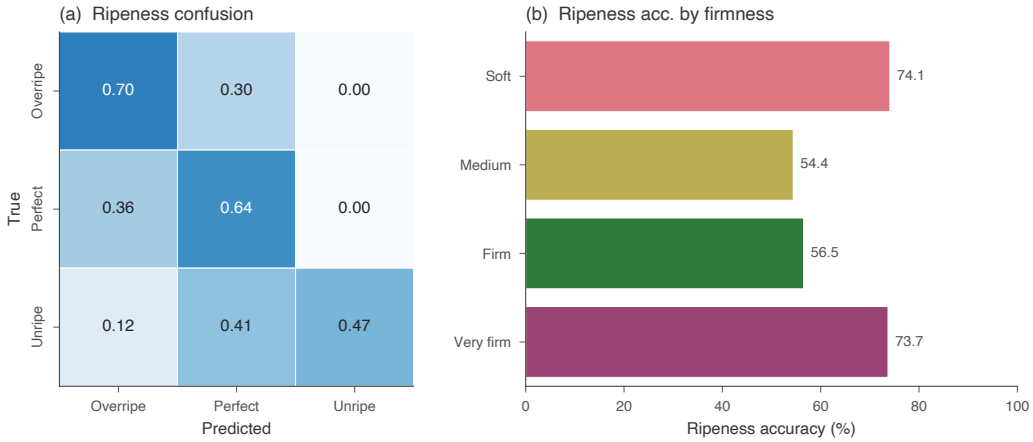

Figure 7: Classification patterns for the best model (ExtraTrees, single-split holdout, n=138). **(a)** Normalized ripeness confusion matrix showing primary confusion between adjacent maturation states. **(b)** Ripeness accuracy stratified by firmness group (4-bin); extreme-firmness groups achieve higher ripeness accuracy than intermediate groups.

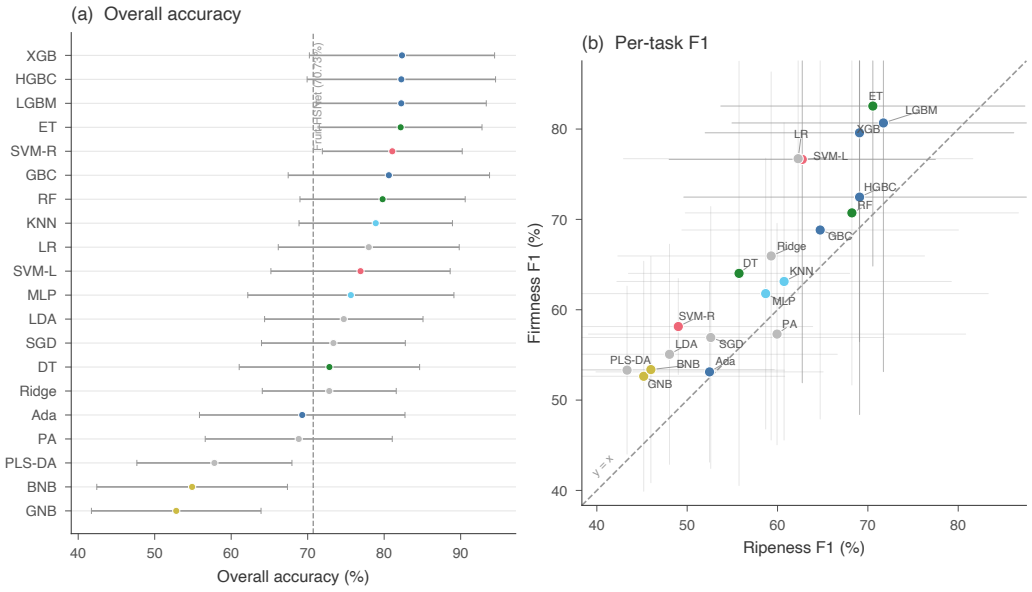

Figure 8: Cross-validation stability (Phase 4, 10-fold stratified CV). **(a)** Overall accuracy with 95% confidence intervals (mean $\pm$ $t_{9,0.025}$·std, where $t_{9,0.025} \approx 2.262$) for all 20 models, sorted by mean. **(b)** Task-separated view: ripeness vs firmness accuracy with CIs. Firmness predictions are consistently more accurate and stable.

Ensembles combined the top five base learners using four aggregation strategies under original unbalanced and optimal preprocessing configurations but did not surpass the best single model in either modality (Table 8, Fig. 9). The best full-spectrum ensemble (hard/soft voting with stratified resplit) achieved 70.65% overall accuracy versus 75.00% for ExtraTrees alone, a 4.35 percentage point degradation.

Table 8: Ensemble performance comparison. All ensembles use the top-5 models from Phase 3. ΔET shows difference from best single model (ExtraTrees, 75.0%).

Strategy	Data State	Ripe↑	Firm↑	OA↑	ΔET
Hard Voting	Original	52.9	85.5	69.2	-5.8
Soft Voting	Original	50.0	84.8	67.4	-7.6
Stacking	Original	31.2	75.4	53.3	-21.7
Blending	Original	49.3	72.5	60.9	-14.1
Hard Voting	Strat. Resplit	57.2	84.1	70.7	-4.3
Soft Voting	Strat. Resplit	58.0	83.3	70.7	-4.3
Stacking	Strat. Resplit	55.8	81.2	68.5	-6.5
Blending	Strat. Resplit	57.2	82.6	69.9	-5.1

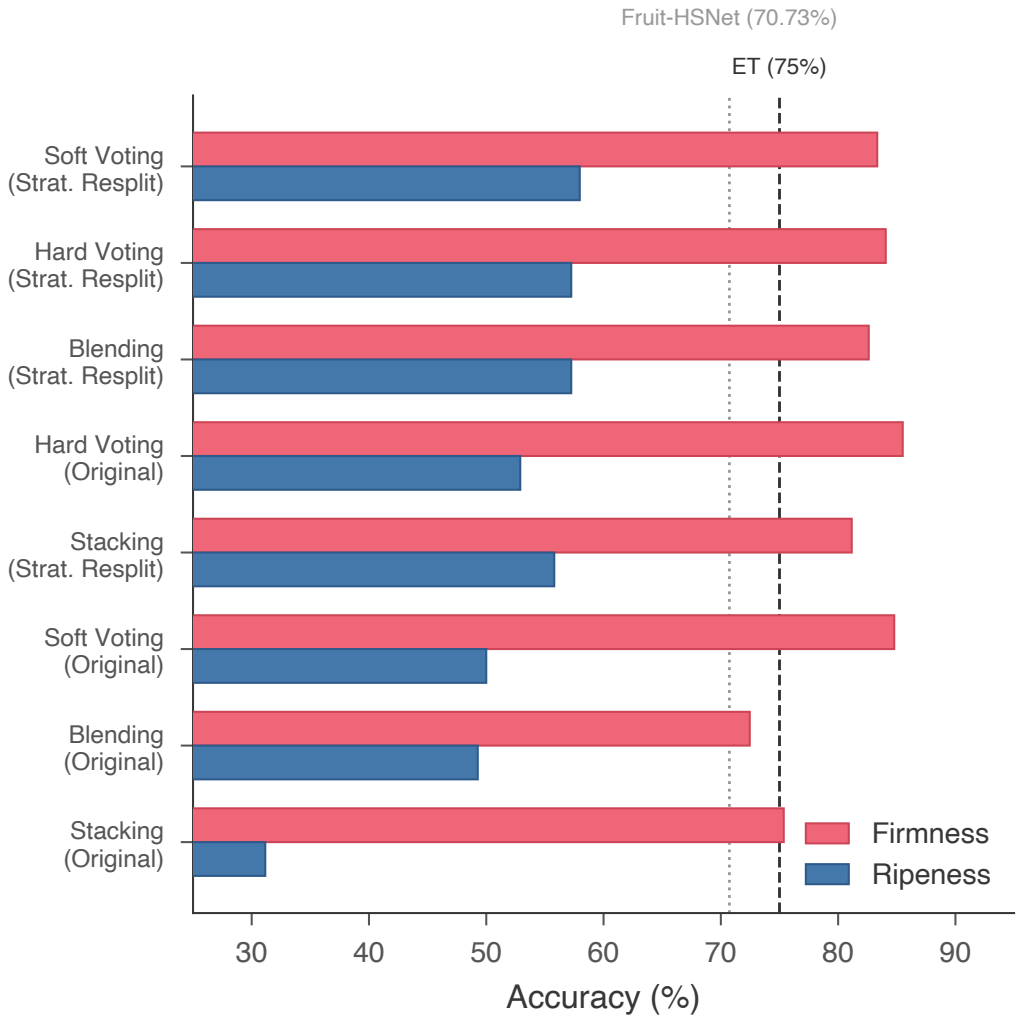

Figure 9: Ensemble performance decomposition by task. All eight ensemble configurations maintain high firmness accuracy (>72%) but catastrophically degrade ripeness accuracy (stacking drops to 31.2%). Dashed line: ExtraTrees single-model baseline (75.0%); dotted line: Fruit-HSNet benchmark (70.73%).

4.5. Explainable AI Analysis

Phase 6 XAI analysis for ExtraTrees revealed concentrated spectral band importance in three regions (Fig. 6c, d): visible blue/green (400–550 nm) associated with pigment changes, red-edge transitions (∼700–750 nm) indicating vegetation stress, and near-infrared (∼770–860 nm) correlated with cellular structure. Feature group ablation (Fig. 10) confirmed that the five spectral transformations contribute differently across tasks and models. For ripeness, first derivatives are the most critical group for LightGBM (8.3% drop when removed), while other preprocessing groups are redundant or harmful for ExtraTrees (removing continuum-removed features improves performance by 10.8%). For firmness, continuum-removed features dominate ExtraTrees (16.6% drop), while all five groups contribute more uniformly to LightGBM (largest drop 5.0%). These task- and model-dependent feature contributions explain why the full 1,120-dimensional feature space outperforms PCA reduction: each transformation captures complementary discriminative information that varies by task.

Ripeness prediction drew primarily on visible-spectrum features associated with chlorophyll degradation and carotenoid accumulation, while firmness prediction relied more heavily on near-infrared features correlated with cellular structure and water content. This wavelength-task correspondence aligns with established plant physiology knowledge (Khodabakhshian and Emadi, 2017; Feng et al., 2023). However, XAI attributions may partly reflect species-level spectral differences rather than within-species quality-state variation. The PCA and t-SNE projections (Fig. 1c,d) show that fruit type dominates the variance structure, so wavelength regions that separate species

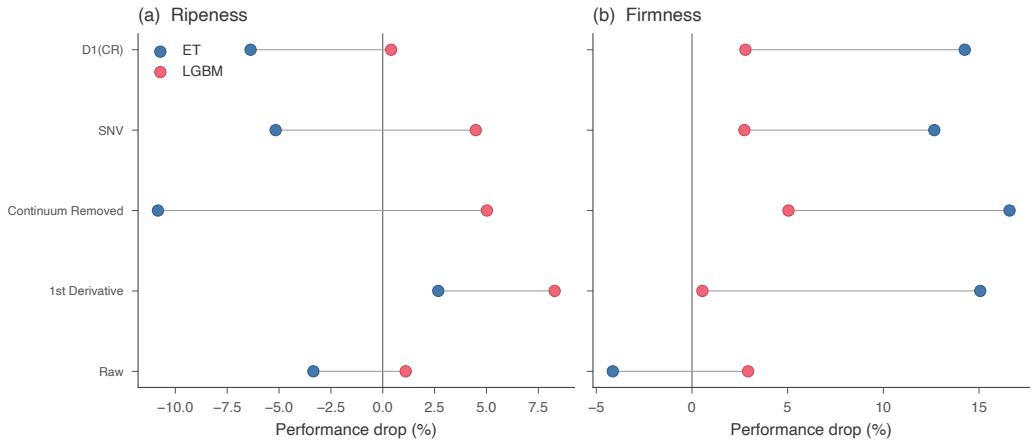

Figure 10: Feature group ablation: relative performance drop (%) when each spectral transformation group (224 bands) is removed. **(a)** Ripeness: first derivatives are most critical for LightGBM (8.3% drop); other groups are redundant or harmful for Extra-Trees. **(b)** Firmness: continuum-removed features dominate ExtraTrees (16.6% drop), while LightGBM shows uniform sensitivity across groups ($\leq$5.0%). Positive values indicate removing the group hurts performance; negative values indicate removal improves it. Task- and model-dependent contributions explain why PCA, which compresses all groups uniformly, degrades classification.

could receive high importance scores even if they carry limited quality information within a single species.

5. Discussion

The comprehensive six-phase evaluation reveals three recurring patterns across 20 algorithms applied to ripeness classification and firmness prediction across five fruit species. Nevertheless, the results raise three fundamental questions that span all evaluated approaches. What mechanism makes preprocessing as influential as algorithm choice? Why does the firmness-ripeness gap persist regardless of model family or modality? And why does PCA, a standard step in hyperspectral workflows, consistently degrade accuracy on this feature space? We discuss each in turn before addressing methodological observations and limitations.

5.1. The Interplay of Preprocessing and Algorithm Choice

The relative importance of preprocessing versus algorithm choice depends on the model family. A bootstrap analysis (10,000 resamples) of the 20×10 Phase 2 accuracy matrix yields a 95% CI of [−5.8, 21.9] pp on the difference between the two median ranges (algorithm: 32.4 pp; preprocessing: 25.2 pp), an interval that contains zero. Neither source of variance can be declared the unambiguous dominant factor with confidence.

For the best-performing models, however, preprocessing is the dominant lever: the best preprocessing configuration improved 15 of 20 models over their Phase 1 baselines, while Phase 3 optimization did not improve any model beyond its best Phase 2 result. For weaker models (LDA, Ridge), preprocessing gains are larger in absolute terms (up to 23 pp) but algorithm

choice matters more; these models remain far below competitive accuracy regardless of preprocessing. Fig. 11 decomposes each model's accuracy trajectory into preprocessing gain (Phase 1→Phase 2) and optimization gain (Phase 2→Phase 3), directly visualizing this asymmetry.

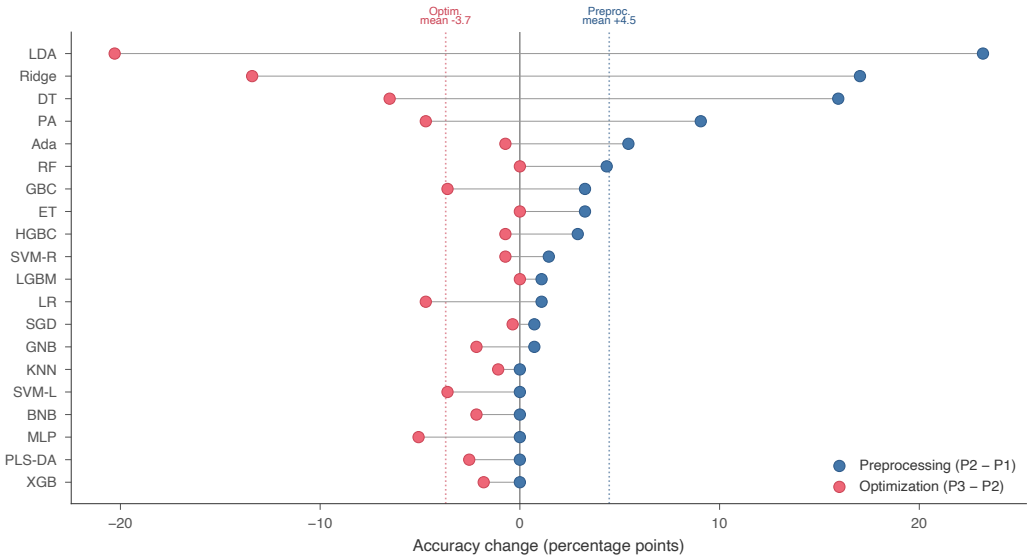

Figure 11: Preprocessing gain vs optimization gain for all 20 models. Each model's accuracy improvement is decomposed into preprocessing gain (best Phase 2 config minus Phase 1 baseline, teal) and optimization gain (Phase 3 minus best Phase 2, purple). Models sorted by preprocessing gain. Dotted lines indicate means. For most models, preprocessing contributes the dominant share of improvement; optimization gain is near-zero or negative for many.

The mechanism behind stratified resplit's advantage over synthetic resampling (SMOTE, random oversampling) likely stems from the structure of spectral data. SMOTE generates new samples by interpolating between existing ones in 1,120-dimensional feature space, an operation that assumes local linearity in spectral manifolds shaped by nonlinear biochemical pro-

cesses. Interpolating between two spectra at different maturation stages does not necessarily produce a spectrum corresponding to an intermediate state. Stratified resplit avoids this problem by redistributing genuine samples across train and test splits without fabrication. However, the stratified resplit condition also increases training set size by 8.7% (414 vs 381 samples) and improves species balance, so its advantage may partly reflect additional data and improved species representation rather than purely a balancing effect.

The practical implication is that for competitive models, preprocessing investment yields larger returns than hyperparameter tuning, while algorithm selection establishes the performance ceiling. Most published studies on this dataset evaluate a single preprocessing pipeline without systematic ablation. Our Phase 2 results across 200 configurations (Fig. 5) demonstrate that varying the preprocessing pipeline alone shifts accuracy by up to 23 pp for the same algorithm (LDA: 35.5% to 58.7%), which may obscure the true source of performance variation in published comparisons.

5.2. Task Asymmetry as a Construct Validity Question

The persistent firmness–ripeness gap is not simply a matter of task difficulty; it reflects a difference in what the two labels measure.

Firmness is a continuous mechanical property with direct spectral correlates. Within the limited NIR range of the FX10 sensor (770–960 nm), absorption features respond to changes in cellular structure and water content associated with tissue softening (Lu et al., 2020; Khodabakhshian and Emadi, 2017). Discretizing firmness into three classes (soft, medium, firm) creates boundaries that align with real physical thresholds. Ripeness, by contrast, is a composite construct. A single ripeness label aggregates chlorophyll

degradation, carotenoid accumulation, sugar-acid balance, volatile production, and textural changes, processes that occur at different rates and may produce conflicting spectral signatures within the same fruit. The confusion between adjacent ripeness states (overripe/perfect: 33.3% mean bidirectional misclassification, unripe→perfect: 41.2%; Fig. 7a) reflects the inherent ambiguity of discretizing a multi-dimensional process into ordinal categories.

The persistence of this gap across all 20 models and both modalities suggests the asymmetry is a property of the labeling scheme rather than of specific algorithms or sensors. Feng et al. (2023) predicted multiple loquat quality parameters and found that prediction accuracy varied across attributes (color $R^2_P = 0.96$, firmness $R^2_P = 0.87$, SSC $R^2_P = 0.84$), consistent with the view that different quality dimensions have inherently different spectral predictability. Future work should investigate whether continuous regression targets or multi-dimensional maturity indices reduce the information loss inherent in categorical ripeness labels.

For deployment, the asymmetry is more acceptable than it first appears. Firmness directly determines shelf life and consumer satisfaction, and commercial distributors prioritize rejecting unacceptably soft fruit over classifying exact ripeness state. However, per-species firmness accuracy (Table 6) remains below 70% for most individual fruit types at Phase 4, which is insufficient for fully autonomous commercial sorting without human oversight. The results establish a useful benchmark but should not be interpreted as a deployable system specification.

5.3. PCA Degradation and the Variance–Discrimination Mismatch

The consistent PCA degradation contradicts standard practice in hyperspectral analysis (Lu et al., 2020; Ahmad et al., 2022). The explanation is rooted in our feature engineering pipeline.

The 1,120-feature space concatenates five spectral transformations, each designed to isolate different physical phenomena: raw reflectance captures absolute intensity, first derivatives capture absorption edge slopes, continuum removal normalizes band depth, and SNV corrects for scattering. PCA maximizes variance across the concatenated space, but variance and task-relevant discrimination are not aligned. First derivatives contribute <0.1% of total variance across the 1,120 features, yet removing them drops LightGBM ripeness accuracy by 8.3% (Fig. 10a). PCA at 95% retained variance effectively discards these low-variance, high-discrimination features.

The 7 models that selected PCA in their optimal Phase 2 configuration were predominantly linear (Ridge, LDA, SGDClassifier, PassiveAggressive). These methods benefit from PCA because dimensionality reduction mitigates the multicollinearity that destabilizes their coefficient estimates. Tree-based methods face no such constraint: their split-based decision boundaries are invariant to feature correlation. This interaction between model family and preprocessing underlines why blanket preprocessing recommendations of always applying PCA to hyperspectral data can be counterproductive when engineered features precede dimensionality reduction.

5.4. Full-Spectrum Versus Reduced-Band Sensing

The full-spectrum advantage at the top of the leaderboard is modest: ExtraTrees achieves 75.0% with 1,120 features compared to 72.1% with ap-

proximate RGB bands (96.1% recovery) and 71.0% with XAI-derived VIS-3 bands (94.7% recovery). Hyperspectral systems cost $10,000–100,000 compared to $500–5,000 for multispectral or RGB alternatives (Min et al., 2023). Both reduced-band configurations recover 94–96% of full-spectrum accuracy using 15 features, and several models achieve comparable or better Phase 3 OA in the low-dimensional space (Table 5).

The VIS-3 and RGB configurations produce statistically indistinguishable results across 20 models (mean OA 61.7% vs 60.1%; Cohen's $d = 0.27$, $p = 0.30$; Table 3). The two band sets differ by only 1–4 spectral indices (2–11 nm in wavelength), yet neither configuration holds a consistent advantage. This insensitivity to exact band placement indicates that the discriminative spectral content in the visible range is spectrally broad: any three well-separated bands within the 400–700 nm window capture it. XAI-guided band selection provides no measurable benefit over generic RGB-center wavelengths on this benchmark.

XAI analysis concentrates discriminative wavelengths in three regions: visible blue/green (400–550 nm) for pigment assessment, the red edge (∼700–750 nm) for chlorophyll transitions, and near-infrared (∼770–860 nm) for cellular structure (Fig. 6c,d). This concentration aligns with the spectral-biochemical relationships established by Khodabakhshian and Emadi (2017) and Feng et al. (2023). However, interpretations specific to the short-NIR boundary (770–960 nm) should be treated as speculative relative to the well-established core NIR firmness bands (1200–1450 nm), which fall outside the Specim FX10 range; the firmness-related importance observed here may partly capture species-level variance rather than within-species quality

differences.

Two caveats apply. First, both reduced-band configurations extract features from hyperspectral cubes at exact wavelengths, bypassing the spectral response curves, Bayer filter artifacts, and noise characteristics of real sensors. The reduced-band accuracy figures represent upper bounds. That said, the RGB parity result weakens the "upper bound" framing: if approximate RGB wavelengths match XAI-optimized bands, the gap between simulated and real sensor performance may be smaller than the VIS-3 result alone suggests. Second, the full-spectrum advantage may increase for tasks requiring finer spectral discrimination, such as early-stage decay detection where subtle NIR absorption shifts precede visible symptoms (Min et al., 2023).

5.5. Tree-Based Dominance and Ensemble Underperformance

Tree-based ensembles (ExtraTrees, RandomForest) occupied the top two positions in Phase 3 and remained competitive in Phase 4 cross-validation ($82.2 \pm 4.7\%$ and $79.8 \pm 4.8\%$). Their advantage arises from a structural match with the feature space. The 1,120 features comprise five transformation groups (224 bands each) with high within-group correlation but distinct cross-group discriminative content (Fig. 10). ExtraTrees' random split-point selection (Geurts et al., 2006) acts as implicit random subspace sampling across these groups, decorrelating individual trees without explicit feature selection. Linear methods face a different problem: multicollinearity among the 224 correlated bands within each transformation inflates coefficient variance and degrades generalization.

The ensemble underperformance with homogeneous base learners (Table 8) is instructive. All four aggregation strategies degraded performance

relative to ExtraTrees alone, with stacking suffering the worst drop (-21.7 pp under original data). The top-five base learners (ExtraTrees, RandomForest, LGBM, HistGradientBoosting, XGBoost) are all recursive partitioning methods that construct decision boundaries through axis-aligned splits. They make correlated errors, particularly on the same borderline ripeness cases (Table 7). Ensemble theory predicts that aggregation improves predictions only when base learner errors are sufficiently uncorrelated. Including a structurally different learner (e.g., SVM or a neural network) as a base learner might restore the diversity that ensemble methods require, though at the cost of interpretability.

ExtraTrees, despite providing feature importance scores through tree-based splitting, is not fully transparent: predictions for individual samples cannot be explained without post-hoc XAI tools. The same limitation applies to all gradient-boosted methods evaluated here. The interpretability advantage of tree-based methods is relative to black-box neural networks, not absolute.

5.6. Spatial Complexity and Mean-Spectrum Representations

ExtraTrees achieves 75.0% on the benchmark split using only per-sample mean spectra, a representation that retains spectral wavelength information but discards all spatial variation across fruit pixels. Fruit-HSNet's dual-branch architecture (Section 3) processes full spatial cubes yet reports 70.73% under a different protocol. Two interpretations are consistent with this outcome. Spatial variation across fruit pixels may carry little additional discriminative signal beyond what the mean spectrum encodes. Alternatively, mean spectrum aggregation may lose useful spatial information, but any advantage

Fruit-HSNet derives from recovering it may be offset by other architectural differences. Distinguishing these interpretations requires ablation within the Fruit-HSNet framework, which the available data cannot support. The practical implication holds regardless of cause. ExtraTrees trains in 0.2 seconds on a consumer CPU, requires no data augmentation, produces a 2.77 MB model, and achieves 2.1 ms per-sample inference latency.

The three discriminative wavelengths identified by explainability analysis (448, 540, 640 nm) were selected as the top joint-ranked visible-range bands from ExtraTrees consensus feature importance across both tasks (Fig. 6c,d, dotted lines). These wavelengths lie within 2–11 nm of standard RGB filter centers (450, 550, 650 nm), and the controlled comparison confirms that this proximity translates to equivalent classification performance: RGB-center bands recover 96.1% of full-spectrum accuracy versus 94.7% for VIS-3 (Table 5). Commodity RGB cameras use broad, overlapping passbands rather than narrow spectral selections, but the insensitivity to exact band placement observed here suggests that such cameras may capture sufficient discriminative content for this task. Validation with actual RGB sensor data remains necessary to confirm this implication, because Bayer-pattern demosaicing, sensor noise, and spectral crosstalk introduce artifacts absent from simulated band selection.

The spectroscopic basis of these feature importance findings carries direct implications for purpose-built inline sensing hardware. Discriminative information concentrated near the chlorophyll absorption trough ($\sim$680 nm), the red-edge transition (700–750 nm), and the tissue water/structure region (770–860 nm) suggests that a targeted multispectral filter array covering these

regions could capture a substantial portion of the discriminative spectral content available in the full 224-band dataset, at a fraction of hyperspectral system cost. Reducing to 6–12 strategically positioned channels would lower per-sample data volume sufficiently to enable inference on embedded controllers without GPU acceleration, a key deployment constraint in high-speed sorting environments. The demonstrated importance of SNV normalization, illustrated by the 12.7% relative performance drop in ExtraTrees firmness upon its removal (Fig. 10b), is not merely a preprocessing detail but a spectroscopically motivated robustness measure against the multiplicative scatter effects of variable surface geometry and path length inherent to production-line acquisition. On a reduced-band sensor, however, computing SNV statistics across only a handful of spectral points becomes poorly conditioned; alternative scatter correction strategies such as external diffuse reference tiles or integrating sphere-based calibration panels would be required to preserve this robustness in practice.

Push-broom line-scan sensors are architecturally well-matched to conveyor belt translation, but rotating or irregularly presented fruit introduces variable illumination geometry that complicates line-scan acquisition. Snapshot multispectral mosaic sensors, which acquire full spatial and spectral information in a single frame, may be better suited to this geometry at the cost of lower spectral resolution, and represent a complementary hardware direction worth evaluating against the wavelength importance findings reported here.

5.7. Single-Split Versus Cross-Validated Evaluation

A pattern in Table 1 warrants attention: several models rank very differently in Phase 3 (single train-test split) versus Phase 4 (10-fold CV). SVC_RBF achieves only 56.9% in Phase 3 but 81.1±4.0% in Phase 4. LogisticRegression drops to 46.7% in Phase 3 while reaching 78.0±5.2% in Phase 4. Conversely, GaussianNB scores 59.1% in Phase 3 but only 52.8 ± 4.9% in Phase 4.

These discrepancies arise because Phase 3 evaluates on a single held-out set, making results sensitive to the specific composition of that split. Phase 4 cross-validation averages over 10 splits and provides confidence intervals. The models most affected (SVMs, LogisticRegression, LDA) are those whose decision boundaries are sensitive to the training sample composition. Tree-based ensembles show smaller discrepancies because bagging and random subspace sampling provide internal averaging that reduces split sensitivity. Even with identical hyperparameters and data splits, stochastic models such as ExtraTrees may vary across re-trainings due to internal randomization (e.g., random subspace selection); the ∼2 pp gap between ExtraTrees' Phase 3 holdout accuracy (75.0%) and Phase 4 holdout re-evaluation (73.2%) reflects this inherent variance rather than a methodological inconsistency.

This observation has a methodological implication for the field: studies reporting accuracy from a single train-test partition may rank algorithms differently than cross-validated evaluations. The Fruit-HSNet benchmark (70.73%) was itself evaluated on a single split (Ben Jmaa et al., 2025), which limits the precision of any direct comparison. Our Phase 4 cross-validated estimates are more reliable for algorithm ranking, though they evaluate a

different quantity (generalization across random partitions) than Phase 3 (performance on the dataset's original held-out set).

5.8. Limitations

Four limitations affect the generalizability of these findings most directly. First, all results derive from a single dataset (DeepHS Fruit) captured with one camera system (Specim FX10) under laboratory conditions across 47 acquisition days. Field environments introduce illumination variation, background clutter, and 15–30% accuracy degradation (Min et al., 2023). Whether the preprocessing-over-algorithm ordering holds under field conditions remains untested. Per-fruit accuracy (33–67% for individual species at Phase 4) falls well below the threshold required for fully autonomous commercial sorting without human oversight; these results should be interpreted as a research benchmark, not a deployment claim.

Second, the comparison to Fruit-HSNet (70.73%) involves a protocol mismatch: Ben Jmaa et al. (2025) used a different train-test partition and evaluation procedure. The Phase 3 result (ExtraTrees, 75.0%; Wilson 95% CI: [67.6%, 81.8%]) and Fruit-HSNet's reported value are not directly comparable. The ExtraTrees confidence interval contains the Fruit-HSNet value. The scikit-learn `MLPClassifier` benchmarked here is a shallow feedforward network, not a deep learning model. ExtraTrees' 75.0% numerically exceeds Fruit-HSNet's 70.73%, but the CI contains that value and protocol differences preclude a controlled comparison; no general superiority claim over deep learning architectures is warranted. The benchmark is scoped to the scikit-learn ecosystem.

Third, sample sizes are imbalanced across fruit species (papaya n=51 vs

avocado n=170), and papaya consistently underperforms (33–44% accuracy; Table 6). Per-fruit deployment would require species-specific calibration and substantially larger training sets for underrepresented varieties.

Fourth, PLS-DA, the standard chemometric baseline, is included but ranks last among the 20 models (P4 CV: $57.8 \pm 4.5\%$), likely because the effective number of latent components is clamped to the number of classes ($n = 3$), severely limiting its capacity on the 1,120-feature space. This result should not be interpreted as evidence that PLS-DA is inherently inferior for spectral classification; with larger class counts or continuous regression targets, PLS-DA may perform competitively.

6. Conclusion

In this study, we investigated various preprocessing strategies and machine learning algorithms for non-destructive estimation of fruit ripeness and firmness from hyperspectral imaging. Our evaluations showed that preprocessing choices can influence hyperspectral fruit quality prediction as strongly as algorithm selection, with preprocessing improvements sometimes exceeding gains from model optimization. The consistent performance gap between firmness and ripeness prediction appears to arise from differences in label structure rather than algorithmic limitations. Tree-based ensemble models achieved up to 82.2% cross-validated accuracy on standard consumer CPU hardware. Additionally, explainable AI analysis identified three informative visible-range wavelengths that retained 94.7% of full-spectrum performance using only 15 features, highlighting their potential for targeted multispectral sensor design. Future work should validate these findings across larger multi-

species datasets to assess the generalizability of the preprocessing–algorithm relationship and the transferability of the identified wavelengths.

Appendix A. Supplementary Figures

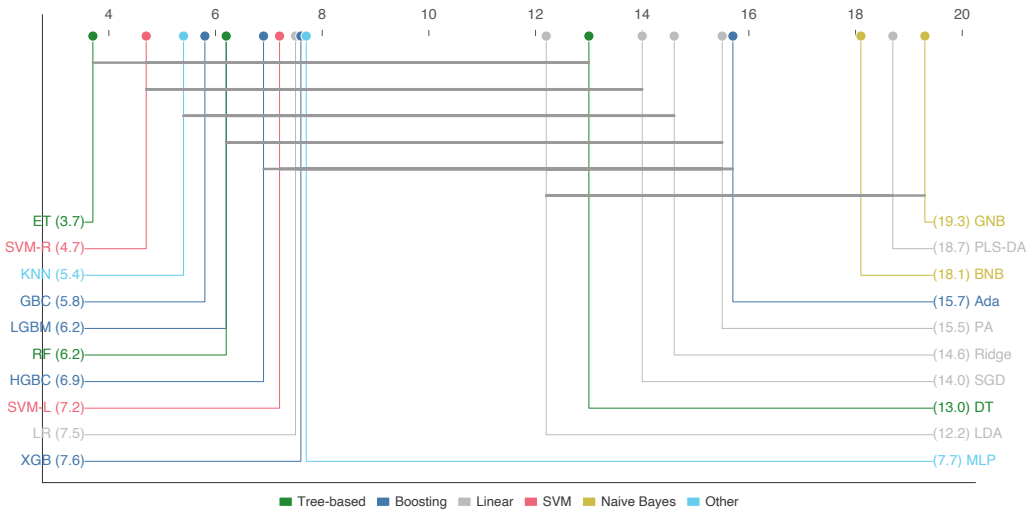

Figure A.1: Critical difference diagram from Nemenyi post-hoc test ($\alpha = 0.05$) on 20-model $\times$ 10-fold overall F1 scores. Models connected by a horizontal bar are not significantly different. Models colored by algorithm family.

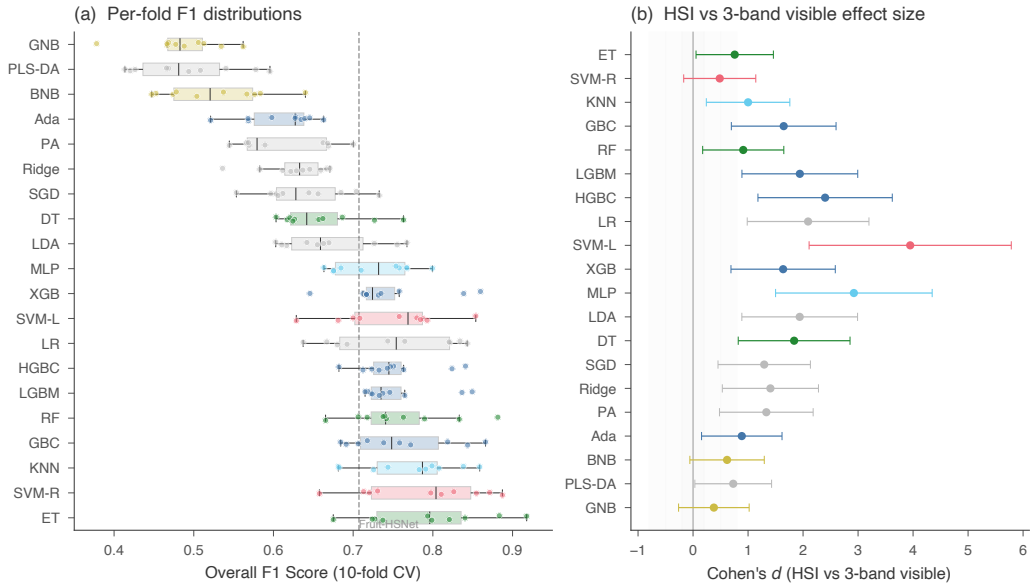

Figure A.2: (a) Per-fold overall F1 score distributions from 10-fold stratified cross-validation, sorted by mean F1. Box plots show median and interquartile range; individual fold scores are overlaid as jittered points. Dashed line marks the Fruit-HSNet benchmark (70.7%). (b) Per-model Cohen's d effect sizes for the HSI vs. VIS-3 subset modality comparison, with 95% confidence intervals. Shaded bands indicate negligible ($|d| < 0.2$), small (0.2–0.5), and medium (0.5–0.8) effect regions. Models colored by algorithm family.

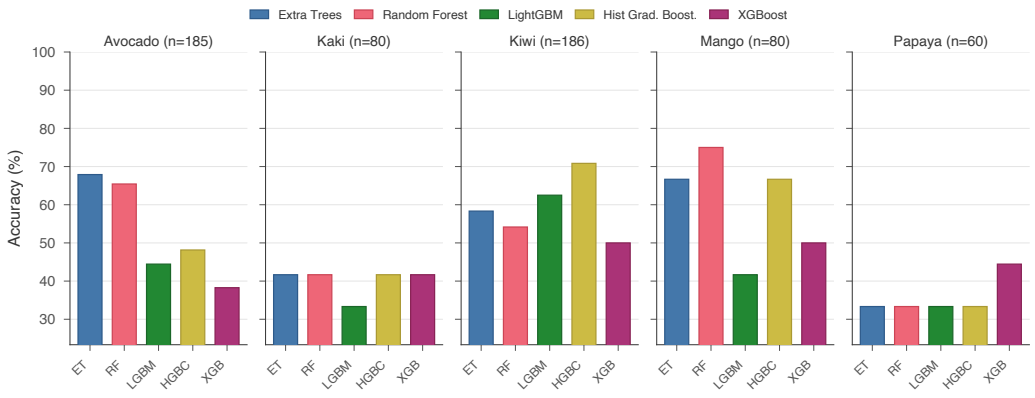

Figure A.3: Per-fruit overall accuracy for the top-five models (Phase 4). Avocado and mango show highest performance; papaya consistently underperforms due to smallest sample size (n=51). Performance variation across species highlights deployment considerations for multi-fruit sorting systems.

Appendix B. Supplementary Tables

Table B.1: Overview of 20 evaluated models with Optuna hyperparameter search spaces. HP = number of tuned hyperparameters; FI = native feature importance. All ranges define the Bayesian optimization bounds (100 trials per model).

Abbr.	Model	HP	FI	Lib.	Optuna Search Space
ET	Extra Trees	5	✓	sklearn	n_est∈[50,500], depth∈[3,30], split∈[2,20], leaf∈[1,20], crit∈{gini,ent}
RF	Random Forest	6	✓	sklearn	n_est∈[50,500], depth∈[3,30], split∈[2,20], leaf∈[1,20], feat∈{sqrt,log2,∅}, crit∈{gini,ent}
DT	Decision Tree	4	✓	sklearn	depth∈[3,30], split∈[2,20], leaf∈[1,20], crit∈{gini,ent}
XGB	XGBoost	9	✓	xgboost	n_est∈[50,500], lr∈[.01,.3], depth∈[3,10], child_w∈[1,10], γ∈[0,1], sub∈[.5,1], col∈[.5,1], α∈[0,1], λ∈[0,1]
LGBM	LightGBM	8	✓	lightgbm	n_est∈[50,500], lr∈[.01,.3], depth∈[3,15], leaves∈[10,300], child∈[5,100], α∈[0,1], λ∈[0,1], col∈[.5,1]
GBC	Gradient Boosting	5	✓	sklearn	n_est∈[50,500], lr∈[.01,.3], depth∈[3,10], split∈[2,20], leaf∈[1,20]
HGBC	HistGradBoosting	5	–	sklearn	lr∈[.01,.3], iter∈[100,500], depth∈[3,10], leaf∈[10,100], L2∈[0,1]
Ada	AdaBoost	3	✓	sklearn	n_est∈[25,200], lr∈[.1,1.5], alg=SAMME
LR	Logistic Regression	3	–	sklearn	C∈[.01,100], pen∈{L1,L2}, solv∈{liblinear,saga}
Ridge	Ridge Classifier	2	–	sklearn	α∈[.1,100], solver∈{auto,svd,cholesky,lsqr,sag,saga}
SGD	SGD Classifier	4	–	sklearn	α∈[1e-6,.01], pen∈{L1,L2}, lr∈{const,opt,inv,adapt}, η_0∈[.01,1]
PA	Passive-Aggressive	3	–	sklearn	C∈[.01,10], iter∈[100,2k], tol∈[1e-5,.01]
LDA	Linear Discriminant	2	–	sklearn	solver∈{svd,lsqr,eigen}, shrink∈{∅,auto,.1,.5,.9}
SVM-L	SVC (Linear)	3	–	sklearn	C∈[.01,100], tol∈[1e-5,.01], iter∈[1k,5k]
SVM-R	SVC (RBF)	3	–	sklearn	C∈[.01,100], tol∈[1e-5,.01], γ∈[1e-6,.1]
GNB	Gaussian NB	1	–	sklearn	var_smooth∈[1e-10,1e-6]
BNB	Bernoulli NB	2	–	sklearn	α∈[.1,10], fit_prior∈{T,F}
KNN	k-NN	4	–	sklearn	k∈[3,50], w∈{uniform,dist}, alg∈{auto,ball,kd,brute}, p∈{1,2}
MLP	MLP	4	–	sklearn	hidden∈{(50),(100),(50,50),(100,50),(100,100)}, α∈[1e-4,.1], lr∈[.001,.1], act∈{relu,tanh,logistic}
PLS-DA	PLS Discriminant Analysis	1	–	sklearn	n_comp∈[1,30] (clamped to $\min(n_{\text{samples}}, n_{\text{features}}, n_{\text{classes}})$)

Table B.2: Best hyperparameters found by Optuna (100 Bayesian TPE trials per model). Parameters listed after stripping the `classifier__` prefix. Models ranked by Phase 3 overall accuracy.

Rk	Model	Best Hyperparameters
1	ExtraTrees	`n_estimators=277, max_depth=14, min_samples_split=6, min_samples_leaf=1, criterion=gini`
2	RandomForest	`n_estimators=285, max_depth=10, min_samples_split=5, min_samples_leaf=2, max_features=log2, criterion=entropy`
3	LGBM	`n_estimators=470, learning_rate=0.2874, max_depth=14, num_leaves=10, min_child_samples=90, reg_alpha=0.2373, reg_lambda=0.4189, colsample_bytree=0.6112`
4	HistGradientBoosting	`learning_rate=0.1795, max_iter=123, max_depth=5, min_samples_leaf=41, l2_regularization=0.5475`
5	XGBoost	`n_estimators=400, learning_rate=0.1357, max_depth=10, min_child_weight=8, gamma=0.1722, subsample=0.7181, colsample_bytree=0.5824, reg_alpha=0.3973, reg_lambda=0.5746`
6	KNeighbors	`n_neighbors=3, weights=distance, algorithm=ball_tree, p=1`
7	GradientBoosting	`n_estimators=181, learning_rate=0.1839, max_depth=8, min_samples_split=11, min_samples_leaf=17`
8	AdaBoost	`n_estimators=198, learning_rate=0.9135, algorithm=SAMME`
9	GaussianNB	`var_smoothing=3.15e-09`
10	DecisionTree	`max_depth=8, min_samples_split=17, min_samples_leaf=8, criterion=entropy`
11	SGDClassifier	`penalty=l2, alpha=5.18e-04, learning_rate=adaptive, eta0=0.3370`
12	BernoulliNB	`alpha=0.1026, fit_prior=True`
13	SVC_RBF	`C=30.96, tol=1.17e-05, gamma=0.0019`
14	MLPClassifier	`hidden_layer_sizes=(100, 50), alpha=0.0255, learning_rate_init=0.0017, activation=relu`
15	PassiveAggressive	`C=3.85, max_iter=369, tol=0.0092`
16	SVC_Linear	`C=0.0880, tol=1.56e-05, max_iter=2560`
17	LogisticRegression	`penalty=l2, C=35.64, solver=liblinear`
18	Ridge	`alpha=93.90, solver=cholesky`
19	LDA	`solver=lsqr, shrinkage=0.1000`
20	PLSDA	`n_components=2`

Table B.3: Hardware and Software Environment Specifications. All 20 models were trained on CPU only; no GPU acceleration was used.

Component	**Specification**
Hardware	
Processor	Apple M4 (10-core CPU: 4 performance + 6 efficiency)
Memory	24 GB unified LPDDR5X
Storage	SSD
GPU	None (CPU-only training)
Software Environment	
Operating System	macOS (arm64)
Python	3.13
scikit-learn	1.7.1
XGBoost	3.0.4
LightGBM	4.6.0
Optuna	4.5.0
imbalanced-learn	0.14.0
SHAP	0.48.0
LIME	0.2.0.1
Reproducibility Settings	
Global random seed	42
Seeded components	Python, NumPy, all classifiers, SMOTE, splits, Optuna TPE sampler
Data split	Dataset-provided VIS benchmark split (73.4/26.6)
Cross-validation	StratifiedKFold, k=10, shuffle=True
Optuna trials	100 per model (TPE sampler, seed=42)
Thread pinning	OMP/OpenBLAS/MKL = 1 thread

Data Availability

DeepHS Fruit dataset (version 2) (Varga et al., 2021) is available at `https://github.com/cogsys-tuebingen/deephs_fruit`. Code and supplementary materials will be made available upon publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT Authorship Contribution Statement

Phongsakon Mark Konrad: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Casper Kunstmann-Olsen:** Supervision, Writing – review & editing. **Jacek Fiutowski:** Supervision, Writing – review & editing. **Serkan Ayvaz:** Conceptualization, Supervision, Writing – review & editing, Project administration.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the authors used Claude Opus 4.6 and Claude Sonnet 4.6 (Anthropic) for writing assistance and code development, and Claude Haiku 4.5 (Anthropic) for routine code formatting tasks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Ahmad, M., Shabbir, S., Roy, S.K., Hong, D., Wu, X., Yao, J., Khan, A.M., Mazzara, M., Distefano, S., Chanussot, J., 2022. Hyperspectral image classification—traditional to deep models: A survey for future prospects. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 968–999. doi:10.1109/JSTARS.2021.3133021.

Ahmed, M.T., Monjur, O., Kamruzzaman, M., 2024a. Deep learning-based hyperspectral image reconstruction for quality assessment of agro-product. Journal of Food Engineering 382, 112223. doi:10.1016/j.jfoodeng.2024.112223.

Ahmed, M.T., Wijewardane, N.K., Lu, Y., Jones, D.S., Kudenov, M., Williams, C., Villordon, A., Kamruzzaman, M., 2024b. Advancing sweetpotato quality assessment with hyperspectral imaging and explainable artificial intelligence. Computers and Electronics in Agriculture 220, 108855. doi:10.1016/j.compag.2024.108855.

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 2623–2631. doi:10.1145/3292500.3330701.

Ben Jmaa, A.B., Chaieb, F., Fabijańska, A., 2025. Fruit-hsnet: A machine learning approach for hyperspectral image-based fruit ripeness prediction, in: Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART), SCITEPRESS. pp. 102–111. doi:10.5220/0013110800003890.

Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization, in: Advances in Neural Information Processing Systems, pp. 2546–2554. doi:10.5555/2986459.2986743.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324.

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, ACM. pp. 785–794. doi:10.1145/2939672.2939785.

Feng, S., Shang, J., Tan, T., Wen, Q., Meng, Q., 2023. Nondestructive quality assessment and maturity classification of loquats based on hyperspectral imaging. Scientific Reports 13, 13189. doi:10.1038/s41598-023-40553-3.

Gao, Z., Shao, Y., Xuan, G., Wang, Y., Liu, Y., Han, X., 2020. Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning. Artificial Intelligence in Agriculture 4, 31–38. doi:10.1016/j.aiia.2020.04.003.

Garillos-Manliguez, C.A., Chiang, J.Y., 2021. Multimodal deep learning and visible-light and hyperspectral imaging for fruit maturity estimation. Sensors 21, 1288. doi:10.3390/s21041288.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine Learning 63, 3–42. doi:10.1007/s10994-006-6226-1.

Guo, Z., Zhang, J., Wang, H., Li, S., Shao, X., Dong, H., Sun, J., Geng, L., Zhang, Q., Guo, Y., Sun, X., Xia, L., Darwish, I.A., 2024. Dual-aspect attention spatial-spectral transformer and hyperspectral imaging: A novel approach to detecting aspergillus flavus contamination in peanut kernels. Postharvest Biology and Technology 213, 112960. doi:10.1016/j.postharvbio.2024.112960.

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, pp. 3149–3157. doi:10.5555/3294996.3295074.

Khodabakhshian, R., Emadi, B., 2017. Application of vis/snir hyperspectral imaging in ripeness classification of pear. International Journal of Food Properties 20, S3149–S3163. doi:10.1080/10942912.2017.1354022.

Lanke, N., Chandak, M., 2025. Recent trends in deep learning and hyperspectral imaging for fruit quality analysis: an overview. Journal of Optics doi:10.1007/s12596-025-02881-7.

Liu, D., Zhang, H., Lv, F., Tao, Y., Zhu, L., 2024. Combining transfer learning and hyperspectral imaging to identify bruises of pears across different bruise types. Journal of Food Science 89, 2597–2610. doi:10.1111/1750-3841.17050.

Lu, B., Dao, P.D., Liu, J., He, Y., Shang, J., 2020. Recent advances of hyperspectral imaging technology and applications in agriculture. Remote Sensing 12, 2659. doi:10.3390/rs12162659.

Lu, Y., Huang, Y., Lu, R., 2017. Innovative hyperspectral imaging-based techniques for quality evaluation of fruits and vegetables: A review. Applied Sciences 7, 189. doi:10.3390/app7020189.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, pp. 4766–4777. doi:10.5555/3295222.3295230.

Min, D., Zhao, J., Bodner, G., Ali, M., Li, F., Zhang, X., Rewald, B., 2023. Early decay detection in fruit by hyperspectral imaging–principles and application potential. Food Control 152, 109830. doi:10.1016/j.foodcont.2023.109830.

Nagasubramanian, K., Jones, S., Sarkar, S., Singh, A.K., Singh, A., Ganapathysubramanian, B., 2018. Hyperspectral band selection using genetic

algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. Plant Methods 14, 86. doi:10.1186/s13007-018-0349-9.

Olisah, C.C., Trewhella, B., Li, B., Smith, M.L., Winstone, B., Whitfield, E.C., Fernández, F.F., Duncalfe, H., 2024. Convolutional neural network ensemble learning for hyperspectral imaging-based blackberry fruit ripeness detection in uncontrolled farm environment. Engineering Applications of Artificial Intelligence 132, 107945. doi:10.1016/j.engappai.2024.107945.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830. doi:10.5555/1953048.2078195.

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, ACM. pp. 1135–1144. doi:10.1145/2939672.2939778.

Schmidt, V., Goyal, K., Joshi, A., Feld, B., Conell, L., et al., 2021. CodeCarbon: Estimate and track carbon emissions from machine learning computing. URL: https://github.com/mlco2/codecarbon, doi:10.5281/zenodo.4658424. software.

Su, Z., Zhang, C., Yan, T., Zhu, J., Zeng, Y., Lu, X., Gao, P., Feng, L., He, L., Fan, L., 2021. Application of hyperspectral imaging for maturity

and soluble solids content determination of strawberry with deep learning approaches. Frontiers in Plant Science 12, 736334. doi:10.3389/fpls.2021.736334.

Varga, L.A., Makowski, J., Zell, A., 2021. Measuring the ripeness of fruit with hyperspectral imaging and deep learning, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–8. doi:10.1109/IJCNN52387.2021.9533728.

Vignati, S., Tugnolo, A., Giovenzana, V., Pampuri, A., Casson, A., Guidetti, R., Beghi, R., 2023. Hyperspectral imaging for fresh-cut fruit and vegetable quality assessment: Basic concepts and applications. Applied Sciences 13, 9740. doi:10.3390/app13179740.

Wieme, J., Mollazade, K., Malounas, I., Zude-Sasse, M., Zhao, M., Gowen, A., Argyropoulos, D., Fountas, S., Van Beek, J., 2022. Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: A review. Biosystems Engineering 222, 156–176. doi:10.1016/j.biosystemseng.2022.07.013.

Yang, C., Guo, Z., Fernandes Barbin, D., Dai, Z., Watson, N., Povey, M., Zou, X., 2025. Hyperspectral imaging and deep learning for quality and safety inspection of fruits and vegetables: A review. Journal of Agricultural and Food Chemistry 73, 10019–10035. doi:10.1021/acs.jafc.4c11492.