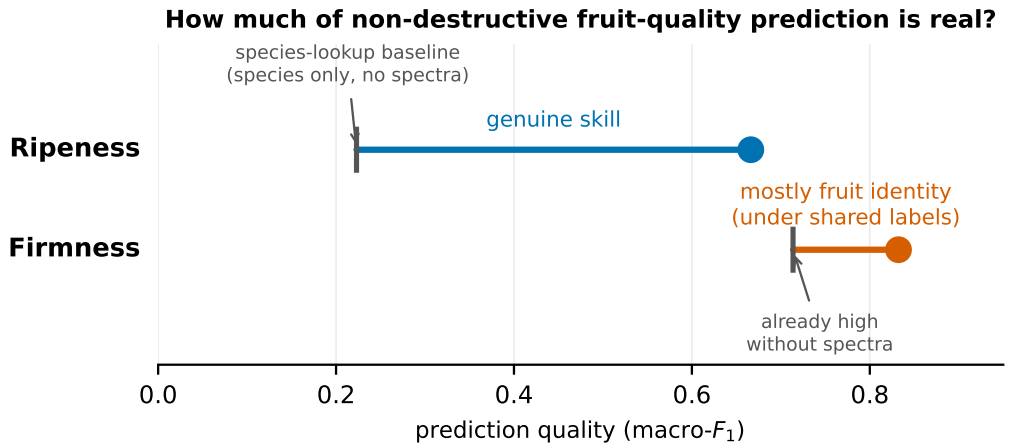
Graphical Abstract

How much of non-destructive fruit-ripeness assessment is real? A confound-controlled, low-cost visible hyperspectral study across five fruit

Phongsakon Mark Konrad, Casper Kunstmann-Olsen, Jacek Fiutowski, Serkan Ayvaz



A baseline using only species separates real spectral skill from fruit recognition.

Highlights

How much of non-destructive fruit-ripeness assessment is real? A confound-controlled, low-cost visible hyperspectral study across five fruit

Phongsakon Mark Konrad, Casper Kunstmann-Olsen, Jacek Fiutowski, Serkan Ayvaz

- A species-lookup baseline separates real spectral skill from fruit recognition
- Visible spectra recover fruit ripeness across five species beyond fruit identity
- The ripeness signal traces to chlorophyll and red-edge pigment bands, within species
- Firmness becomes recoverable under fruit-specific rather than shared labels
- A low-cost visible sensor is a trustworthy, inexpensive route to ripeness grading

How much of non-destructive fruit-ripeness assessment is real? A confound-controlled, low-cost visible hyperspectral study across five fruit

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

Non-destructive optical sensing promises to grade fruit ripeness on the packing line without cutting the fruit open, and public hyperspectral benchmarks report high accuracy across several fruit at once. Pooling fruit species this way hides a problem, because species is the loudest signal in a spectrum and a model can score well by first recognising the fruit. We separate genuine quality prediction from fruit recognition with a baseline that uses species identity and no spectra, and we find that ripeness clears this baseline by a wide margin while a firmness score built from labels shared across species clears it only narrowly. The ripeness signal a low-cost visible sensor recovers traces to the pigment changes of ripening, and it holds up within a single species and under stricter validation. Across five fruit, honest evaluation points to visible-range sensing of ripeness as a trustworthy and inexpensive target for postharvest grading.

Keywords: Hyperspectral imaging, Ripeness, Firmness, Non-destructive quality assessment, Confounding, Model evaluation

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

1. Introduction

A packing house handles tens of thousands of fruit a day and cuts open only a handful to measure quality. Everything else is graded by eye, which reads surface colour and misses the internal state that decides shelf life and sorting. Poor sorting sends unripe fruit to the shelf and lets soft fruit spoil in transit, and a large part of fresh produce is lost after harvest for want of a fast internal check (Porat et al., 2018). Optical sensing offers a way to close this gap. A camera that records reflectance across many wavelengths sees pigment and tissue changes that track ripening, and it does so without touching or destroying the fruit (Lu et al., 2020a,b). Non-destructive visible and near-infrared sensing is by now an established postharvest tool for reading ripeness and firmness in stored fruit (Lee et al., 2025; Ding et al., 2024; Li et al., 2024), and public benchmarks built on the same idea report high accuracy for fruit quality, which reads as a solved problem waiting for hardware.

The reported accuracy is easy to trust and easy to misread. Most benchmarks pool several fruit species and summarise a model with a single number (Varga et al., 2021; Ben Jmaa et al., 2025), and they read one held-out split as sufficient. Yet species is by far the strongest signal in a spectrum, stronger than any single quality state, so a model that pools species can score well by first recognising the fruit, a shortcut that inflates a benchmark score without measuring quality (Geirhos et al., 2020). These benchmarks do not subtract the accuracy a model reaches by identifying the species, so a score that reflects fruit recognition is read as a measured quality trait. The chemometrics literature has begun to document how ordinary evaluation choices, such as splitting and preprocessing selection, leak information and inflate reported performance in spectral models (Király and Tóth, 2025; Ezenarro and Schorn-García, 2025). That work establishes that inflation happens in principle, yet the question raised by multi-species pooling stays open. Nobody asks how much of a pooled quality score reflects the quality of the fruit and how much reflects knowing which fruit it is.

A parallel line of work assigns physiological meaning to informative wavelengths and studies reduced-band sensing. Studies relate pigment and water absorption features to ripening and quality (Khodabakhshian and Emadi, 2017; Feng et al., 2023), building on a long history that ties visible reflectance to chlorophyll loss and the red edge during ripening (Merzlyak et al., 1999) and near-infrared bands to firmness (Nicolaï et al., 2007), and others show that a handful of bands can retain classification accuracy (Nagasubramanian

et al., 2018; Zou et al., 2010). These works read band importance on pooled data and assume that the selected bands transfer to cheap sensors. On a set where species dominates the spectral variance, pooled importance can reflect what separates species rather than what separates quality states within a fruit, and whether that importance survives a within-species control has not been tested.

We ask these questions with a baseline that uses species identity and no spectra at all, so that any real spectral skill has to beat it. Read this way, ripeness clears the baseline by a wide margin. Its signal survives a within-species control and stricter validation, and it traces to the pigment bands that change as fruit ripens. Because those bands sit in the visible range, a low-cost visible sensor is enough to recover ripeness, which is a practical result for postharvest grading. Firmness is more delicate. A firmness score built from one labelling scheme shared across species clears the baseline only narrowly, because that scheme lets a hard fruit and a soft fruit map cleanly onto species. Under each fruit's own firmness scale the spectra recover firmness, so the shortfall is a property of cross-species labelling rather than of the trait.

This paper makes three contributions. The first is a species-lookup baseline, paired with a within-species analysis, that separates genuine spectral prediction from fruit recognition in pooled fruit-quality benchmarks. The second is evidence across five fruit that visible spectra recover ripeness through the pigments of ripening, a result that holds under a within-species control and under folds grouped by physical fruit. The third is a demonstration that the weak pooled firmness result is an artefact of cross-species labelling rather than a sensor limit, together with a low-cost visible-sensor account of ripeness in which band count matters more than band placement.

2. Materials and methods

2.1. Data

We use the visible-range subset of the DeepHS Fruit dataset, version 2 (Varga et al., 2021), recorded with a Specim FX10 camera over 224 bands spanning 398 to 1004 nm. The dataset covers five fruit, namely avocado, kiwi, mango, kaki, and papaya, imaged over 47 acquisition days, so the set spans a range of ripening stages. This camera reaches only the short near-infrared, so the water absorption bands past 1200 nm that firmness assessment often relies on are absent. Each sample carries quality labels from destructive measurement. Ripeness is a three-class scheme, unripe, perfect, and overripe,

from visual assessment and destructive sampling, and the source dataset does not report inter-annotator agreement. Firmness is a penetrometer force in grams-force, which the benchmark discretises into soft, medium, and firm classes at thresholds drawn from the pooled cross-fruit distribution rather than from any one fruit. This shared cut is the labelling choice we return to in Section 3.3. We represent every sample by the mean reflectance spectrum of its fruit pixels, a 224-dimensional vector, and use no spatial or engineered features. After removing samples without a valid label the pooled set holds 591 samples for ripeness and 573 for firmness. Figure 1 summarises the class balance, the mean spectra, and a principal component projection of the samples.

2.2. Leakage-controlled evaluation

The prior benchmark selects a preprocessing pipeline on a fixed test split, which lets the test data influence a modelling choice. Spectral pipelines choose among scatter-correction and smoothing transforms such as the standard normal variate, multiplicative scatter correction, and Savitzky-Golay filtering (Barnes et al., 1989; Geladi et al., 1985; Savitzky and Golay, 1964), and selecting any of them on the data used to score leaks information into the estimate. We remove this by pooling all samples and scoring with nested cross-validation (Cawley and Talbot, 2010; Varma and Simon, 2006). An outer loop of ten stratified folds scores only, and an inner loop of five folds selects preprocessing and hyperparameters on the training part of each outer fold. No preprocessing decision sees the fold it is scored on.

2.3. Evaluation metrics

Every model is scored with the macro-averaged F_1, which weights the three classes equally and does not reward a model for following the majority class. For class c with precision P_c and recall R_c, the per-class score and its macro-average across the C classes are

$$F_{1,c} = \frac{2P_cR_c}{P_c + R_c}, \qquad \mathrm{F1}_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C} F_{1,c}. \tag{1}$$

Alongside it we report balanced accuracy, the mean of the per-class recalls,

$$\mathrm{bal.acc} = \frac{1}{C}\sum_{c=1}^{C} R_c, \tag{2}$$

which is likewise not inflated by a large majority class. To quantify how strongly a quality label is tied to species we use Cramér's V,

$$V = \sqrt{\frac{\chi^2}{n\,(k-1)}}, \tag{3}$$

where χ^2 is the chi-square statistic of the species-by-label contingency table, n is the number of samples, and k is the smaller of the species and class counts. The value runs from zero when species and label are independent to one when species fixes the label.

For uncertainty we report 95% confidence intervals obtained by bootstrapping across the outer folds. To check that the results do not hinge on one random partition we repeat the reference evaluation across five seeds and report the spread. To test whether a model's skill over the baseline is real rather than noise we pair each outer fold's model and baseline scores and apply a Wilcoxon signed-rank test, with a percentile bootstrap confidence interval on the mean skill.

2.4. The species-lookup baseline

The central control is a predictor that uses species and no spectra. For a task with labels y and species s, the baseline assigns to every sample of species s the most common label of that species in the training data (Equation 4).

$$\hat{y}_{\text{base}}(s) = \arg\max_{c} \, |\{i : s_i = s,\ y_i = c\}|. \tag{4}$$

For example, if most avocados in the training data are labelled perfect, the baseline calls every avocado perfect and never reads its spectrum. It is a lookup table from species to that species' most common class. We define the spectral skill of a model as its macro-F_1 minus that of this baseline (Equation 5).

$$\text{skill} = \text{F1}^{\text{model}}_{\text{macro}} - \text{F1}^{\text{base}}_{\text{macro}}. \tag{5}$$

A model with high accuracy but low skill is recovering fruit identity, not quality. We also measure how strongly each quality label is tied to species with Cramér's V (Equation 3). Two further checks probe the shortcut directly. The first measures how accurately a classifier recovers species from the spectrum alone, which shows how available fruit identity is, and the second measures how often each model returns the same class the species-lookup baseline would, which shows how far a model reduces to fruit recognition.

2.5. Within-species analysis

Skill above a pooled baseline can still hide a confound, so we repeat the prediction inside each species. For a single species we compare a cross-validated model against the within-species majority-class rate and report the accuracy headroom the model gains over it. A trait that varies within a species leaves headroom to recover. A trait that is nearly constant within a species leaves almost none, regardless of the sensor.

2.6. Reduced-band and sensor-response analysis

To test low-cost sensing we reduce the full spectrum to three bands in two ways, one set chosen from importance analysis in the visible range and one set placed at the centres of a standard colour camera. We compare both against the distribution of macro-F_1 from one thousand random three-band subsets, which fixes the count at three and randomises placement. To move from wavelength slices toward a real sensor we convolve the spectrum with Gaussian passbands of increasing width and add photon noise and illumination variation, then re-score.

2.7. Wavelength importance

For the model with the strongest ripeness skill we rank band importance by permutation (Breiman, 2001; Fisher et al., 2019), globally and within the two largest species. Comparing the two tells us whether the bands a pooled model relies on are the same bands that carry quality within a single fruit.

2.8. Models and robustness

We benchmark twenty classifiers spanning linear, kernel (Cortes and Vapnik, 1995), neighbour, tree ensemble (Breiman, 2001; Geurts et al., 2006), boosting (Chen and Guestrin, 2016; Ke et al., 2017), discriminant, and naive Bayes families, implemented in scikit-learn (Pedregosa et al., 2011) with the Extra Trees model as the reference. Appendix Appendix A lists every model with its hyperparameter search space. A partial least squares discriminant model (Wold et al., 2001; Barker and Rayens, 2003) is included with its latent dimension tuned rather than fixed, because a fixed low setting understates it. Each fruit is imaged from two sides, so a plain shuffle can place both views of the same physical fruit in different folds. To measure this leak we rebuild each physical fruit's identity from its recording file name and repeat the reference evaluation with folds grouped so that a fruit's two views always

share a fold. Appendix Appendix C records the software environment and the full cross-validation configuration.

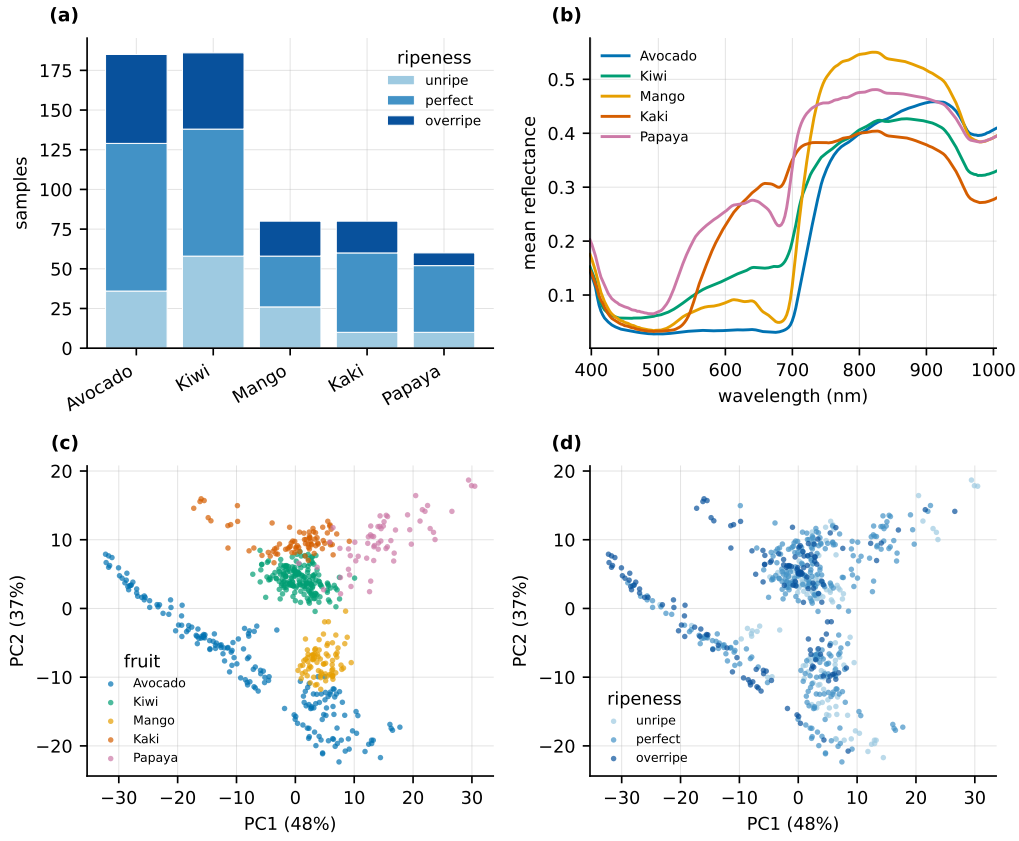


Figure 1: The pooled visible-range dataset. (a) Ripeness class counts per fruit. (b) Mean reflectance spectrum of each fruit. (c) Principal component projection of the spectra coloured by fruit, where the fruit form tight, well-separated clusters. (d) The same projection coloured by ripeness class, where the classes spread across the space without a clean boundary. Species dominates the spectral variance, which is why a model can score on quality by first recognising the fruit and why ripeness is best read within a species.

3. Results

3.1. Reported accuracy hides fruit recognition

Table 1 lists the strongest models by pooled cross-validated macro-F_1 (Equation 1), and Appendix Appendix B gives all twenty. Read on its own the table tells the usual story, with firmness scoring well above ripeness and tree ensembles at the top. Table 2 changes the reading. Ripeness is nearly independent of species, with a Cramér's V of 0.16, so the species-lookup baseline sits near chance and every model adds a wide margin of genuine skill. Firmness is different under the labelling used here. With one soft, medium, and firm cut shared across all five fruit, firmness is tied to species with a Cramér's V of 0.61, and the baseline already reaches a macro-F_1 of 0.71. The best model adds 0.12 over it and the median model adds 0.08, so under this shared scheme the spectra contribute a small margin on top of fruit recognition. Section 3.3 shows that this shortfall is a property of the shared labelling, not of firmness. Figure 2 shows the pattern for every model. On ripeness the whole field sits far to the right of the baseline. On firmness under the shared scheme the field clusters against the baseline line, and the weakest models fall below it.

Two checks tie this pattern to species identity directly. A projection of the spectra already shows the imbalance, with the fruit forming clear clusters while the ripeness classes spread across the same space without a boundary (Figure 1c, d). Species is almost perfectly recoverable from the spectrum, with the reference model classifying the five fruit at 0.99 accuracy, so fruit identity is available to any model as a shortcut. The models take that shortcut on firmness far more than on ripeness. The reference model agrees with the species-lookup baseline, meaning it returns the same class the lookup would, on about three quarters of firmness samples and only about half of ripeness samples. The gap is not specific to one model. Across all twenty models the lowest firmness agreement still exceeds the highest ripeness agreement, so every model behaves more like the species lookup on firmness than on ripeness. A firmness model largely reproduces fruit recognition, whereas a ripeness model departs from it and reads the spectrum.

3.2. The traits separate again within a species

Skill above a pooled baseline could still ride on a residual confound, so we predict inside each species. Table 4 and Figure 3 report the accuracy a model gains over the within-species majority. Ripeness leaves real headroom

Table 1: Top ten models by pooled nested cross-validation macro-F_1 (mean with 95% confidence interval across outer folds), for ripeness and firmness. Higher is better, so the best value in each column is in bold and the second best is underlined. The leading models' intervals overlap, so these marks denote rank rather than a significant separation. Preprocessing is selected inside the inner folds, so the outer folds score only. The ranking read on its own overstates firmness, which Table 2 corrects.

	Ripeness		Firmness	
Rk	Model	$F_1 \uparrow$ [95% CI]	Model	$F_1 \uparrow$ [95% CI]
1	Extra Trees	**0.666** [0.63, 0.70]	Extra Trees	**0.832** [0.79, 0.87]
2	Random Forest	<u>0.653</u> [0.63, 0.68]	PLS-DA	<u>0.828</u> [0.80, 0.86]
3	LightGBM	0.652 [0.63, 0.68]	SVC (RBF)	0.824 [0.79, 0.86]
4	HistGradientBoosting	0.629 [0.60, 0.65]	Random Forest	0.824 [0.78, 0.86]
5	Gradient Boosting	0.628 [0.60, 0.65]	k-NN	0.813 [0.76, 0.86]
6	SVC (RBF)	0.626 [0.60, 0.65]	LightGBM	0.813 [0.77, 0.86]
7	XGBoost	0.625 [0.61, 0.64]	HistGradientBoosting	0.806 [0.76, 0.84]
8	k-NN	0.612 [0.56, 0.66]	Ridge	0.805 [0.77, 0.84]
9	LDA	0.610 [0.58, 0.64]	XGBoost	0.796 [0.74, 0.84]
10	Ridge	0.599 [0.56, 0.64]	Gradient Boosting	0.796 [0.75, 0.84]

in the three larger species, and the gain is largest in avocado. Firmness is different across species. In avocado, where firmness varies, a model gains headroom, so firmness is not unmeasurable in principle. In mango almost every sample falls in one firmness class, the majority rate is already high, and there is nothing left to recover. The pooled firmness score hides this by borrowing the easy cross-species contrast.

3.3. Firmness predictability turns on how firmness is labelled

The soft, medium, and firm classes above use one force cut shared across all five fruit, which lets a naturally hard fruit and a naturally soft fruit fall into different classes by species. Table 3 compares this shared scheme against two fruit-specific alternatives, each fruit's own firmness split into three equal parts and the thresholds the dataset defines for avocado and kiwi. Under the shared scheme a model gains almost nothing over the within-species majority for firm fruit such as mango, because the class is nearly constant within that species. Under fruit-specific labels the same spectra recover firmness with a clear margin in every fruit tested, including gains over the majority of up to about two fifths in accuracy. Firmness is therefore recoverable from the visible spectrum when it is labelled on each fruit's own scale. What the

Table 2: Prediction skill above a species-lookup baseline. The baseline assigns each fruit its most common class and uses no spectra. Cramér's V measures how strongly the quality label is tied to species. Skill is the macro-F_1 of the model minus the macro-F_1 of the baseline. A large skill means the spectra carry information beyond fruit identity. An upward arrow marks a column where a higher value is better, and the larger skill in each of the last two columns is in bold.

Task	V(species)	Baseline F_1	Best model	Model F_1	Best skill ↑	Median skill ↑
Ripeness	0.16	0.22	Extra Trees	0.67	**+0.44**	**+0.37**
Firmness	0.61	0.71	Extra Trees	0.83	+0.12	+0.08

Table 3: Firmness prediction skill (accuracy gained over the within-species majority) under three labelling schemes. The shared scheme uses one soft, medium, and firm cut for all fruit. The within-fruit scheme splits each fruit's own firmness into three equal parts. The dataset scheme uses the thresholds the source dataset defines, which exist only for avocado and kiwi. A dash marks a fruit the dataset does not threshold. More skill is better, so for each fruit the best scheme is in bold and the second best is underlined. Under fruit-specific labels the spectra recover firmness in every fruit, so the weak firmness result under the shared scheme is a labelling effect.

Fruit	Shared cut ↑	Within-fruit thirds ↑	Dataset thresholds ↑
Avocado	+0.20	**+0.42**	+0.23
Kiwi	+0.10	**+0.26**	+0.16
Mango	+0.00	**+0.17**	–
Kaki	+0.06	**+0.14**	–
Papaya	+0.15	**+0.38**	–

pooled benchmark reports as weak firmness prediction is a consequence of harmonising the firmness label across species, not a limit of the sensor.

3.4. The ripeness signal is pigment chemistry

Because ripeness is the trait a spectral sensor recovers, we ask which wavelengths carry it. Figure 4 compares band importance for ripeness on pooled data against importance within avocado and within kiwi. The pooled ranking spreads across the blue and green, a mixture of pigment and species contrast. The within-species rankings sharpen onto the red and the red edge. In avocado the top bands cluster near chlorophyll-a absorption around 670 nm and the red edge near 700 to 720 nm, the region that shifts as chlorophyll breaks down during ripening. In kiwi the importance spans a carotenoid band in the blue and the same red edge. Firmness importance, by contrast, does not

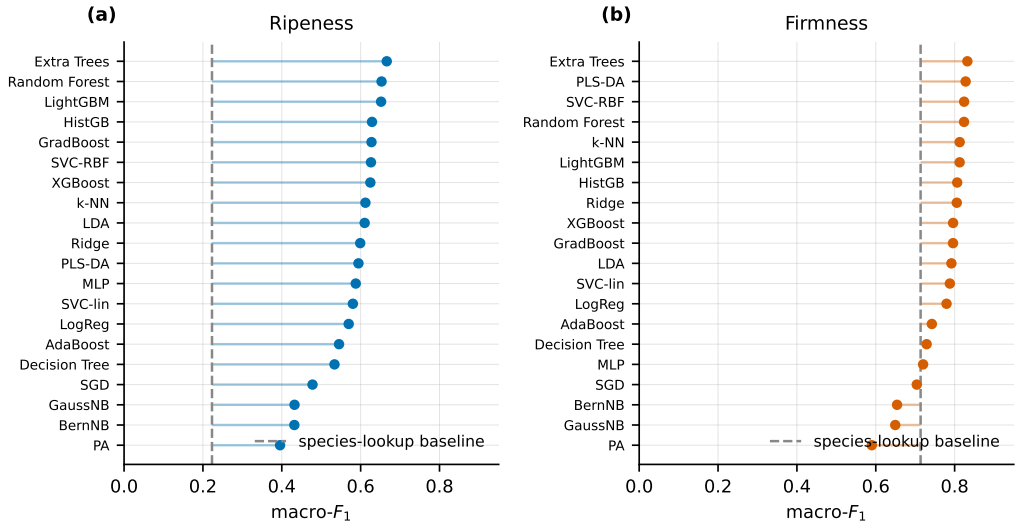


Figure 2: Each model against the species-lookup baseline (dashed), which predicts from fruit identity alone. On ripeness (left) every model clears the baseline by a wide margin, so the spectra carry real information. On firmness (right), under a firmness scheme shared across species, models sit against the baseline and the weakest fall below it. Section 3.3 shows fruit-specific labels change this.

settle on any coherent region once species is controlled, consistent with the small firmness skill in Figure 2.

3.5. Low-cost sensing recovers ripeness, with honest limits

The ripeness signal sits in the visible range, which a low-cost sensor can reach. We test how far the spectrum can be reduced. Figure 5 compares a three-band set chosen by importance against a three-band set placed at colour-camera centres, both against the spread of one thousand random three-band subsets. The chosen and the generic sets land together, and both sit near the random median. Placement barely matters once the count is fixed at three, so the value of the reduced spectrum is its count of well-separated bands rather than any precise choice. This tempers a common reading of importance-guided band selection. Figure 6 then moves from clean wavelength slices toward a real sensor. Widening the passband alone costs little, but adding photon noise and illumination variation pulls accuracy down sharply, most of all for ripeness. A three-band claim measured on clean slices is an upper bound, not a sensor specification.

Table 4: Within-species prediction. For each species we report the within-species majority-class rate and the accuracy headroom a model gains over it (cross-validated Extra Trees). More headroom is better, so the best value in each headroom column is in bold and the second best is underlined. Ripeness gains real headroom in the larger species. Firmness gains headroom only in avocado, where firmness varies, and none in mango, where almost every sample shares one class. A dash marks a fruit whose within-species firmness could not be cross-validated, because a firmness class is too small.

		Ripeness		Firmness	
Species	n	Majority	Headroom ↑	Majority	Headroom ↑
Avocado	185	0.50	**+0.30**	0.72	**+0.20**
Kiwi	186	0.43	+0.17	0.70	<u>+0.10</u>
Mango	80	0.40	<u>+0.23</u>	0.93	+0.00
Kaki	80	0.62	+0.04	0.55	+0.06
Papaya	60	0.70	+0.07	0.57	–

3.6. Robustness

The reference scores hold when preprocessing is chosen inside the inner folds, so the correction is not an artefact of a leaky pipeline. The partial least squares model, tuned rather than fixed, is competitive on firmness rather than last, which shows the earlier last-place reading was a wrapper setting and not a property of the method. Grouping folds by physical fruit, so that the two views of one fruit never split across a fold, lowers the reference macro-F_1 from 0.67 to 0.62 on ripeness and from 0.83 to 0.80 on firmness. The two sides of a fruit are mildly correlated, so a plain shuffle inflates the estimate by three to four points. This drop leaves the conclusions unchanged, because grouped ripeness still clears the species-lookup baseline by a wide margin and grouped firmness under the shared labels still adds little over it.

The pattern also does not depend on the particular five-fruit mix. Across all sixteen three-, four-, and five-fruit subsets, firmness is tied to species more strongly than ripeness in every subset, with a median Cramér's V of 0.61 against 0.15, and every subset leaves more spectral skill for ripeness than for firmness. The confound is therefore a property of pooling species rather than of this specific set of fruit.

The reference result is also stable across seeds. Repeating the nested evaluation over five random partitions moves the reference macro-F_1 by under a hundredth on either task, so the ranking does not rest on one lucky split. A paired Wilcoxon signed-rank test over folds and seeds places both skills

above zero, yet the effect sizes are far apart, a ripeness skill of 0.45 (95% CI 0.43 to 0.47) against a firmness skill of only 0.12 (0.10 to 0.13) under the shared labels. Because cross-validation folds share training data, we read the test as evidence of direction and rely on the effect sizes and the seed spread for magnitude. Firmness adds a small though detectable margin over fruit recognition, and ripeness adds a large one.

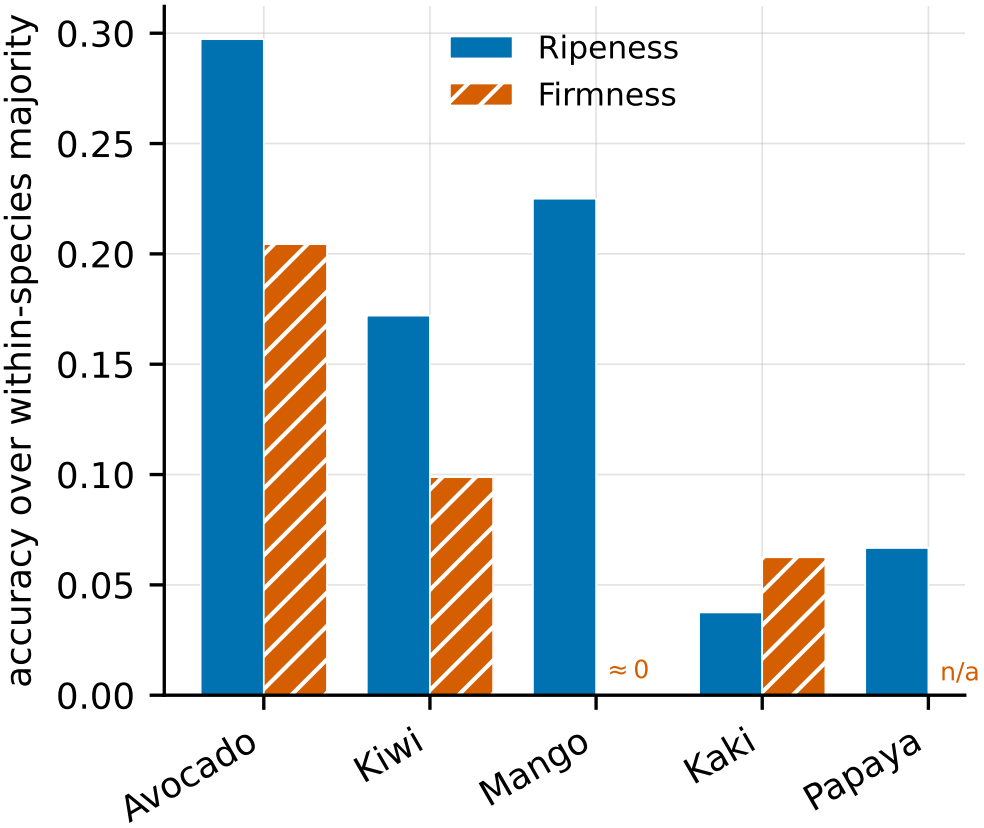


Figure 3: Accuracy a model gains over the within-species majority. Ripeness leaves headroom to recover in the larger species. Firmness leaves headroom only in avocado and near zero in mango, where the trait is nearly constant. Papaya firmness is marked n/a because a within-species firmness class is too small to cross-validate.

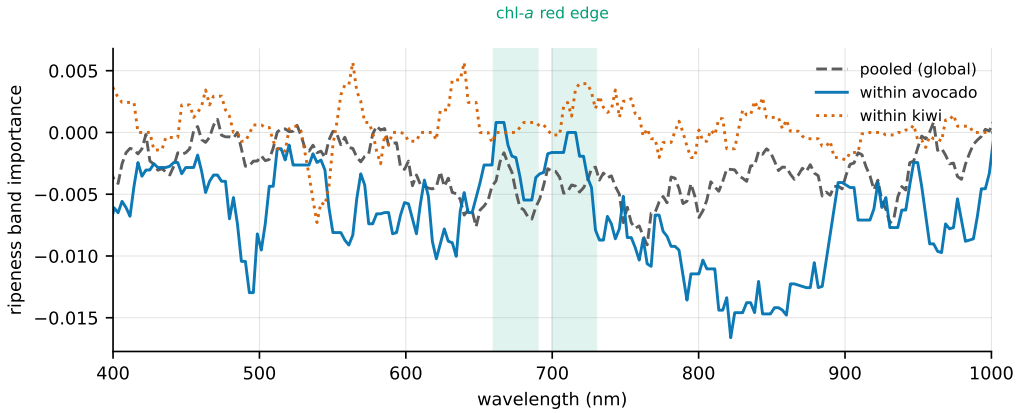


Figure 4: Ripeness band importance, pooled and within two species. Pooling spreads importance across the visible range. The within-species control sharpens it onto chlorophyll and red-edge bands (shaded), the pigment transitions of ripening.

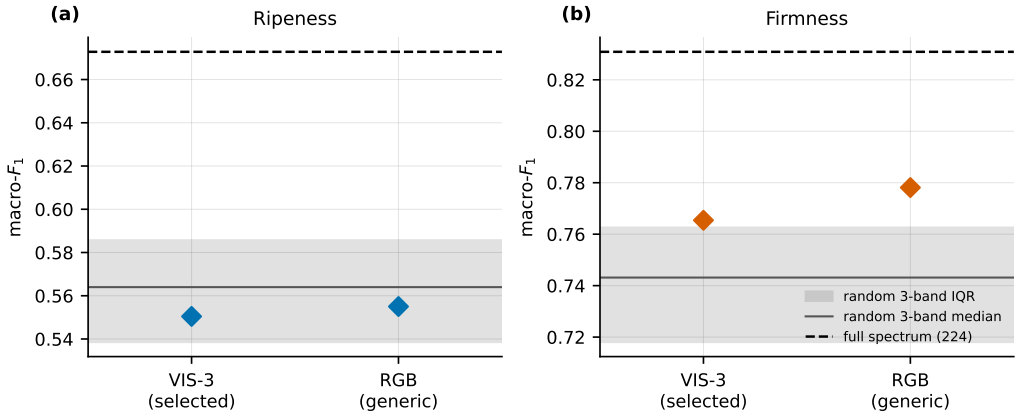


Figure 5: Three-band accuracy against the spread of random three-band subsets and the full spectrum. The importance-chosen and colour-camera sets land together near the random median, so band count rather than placement drives the reduced-band result.

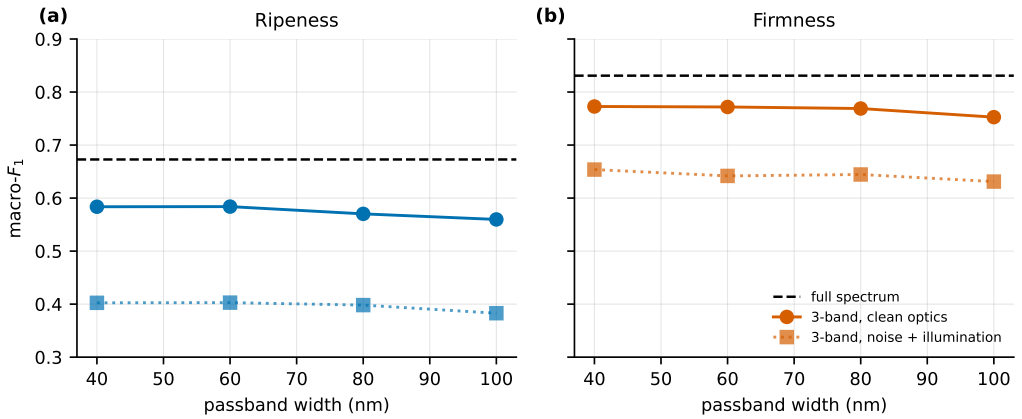


Figure 6: Reduced-band accuracy under a modelled sensor. A wider passband alone costs little, but noise and illumination variation erode accuracy, so the clean three-band figure is an upper bound.

4. Discussion

The result that matters for practice is that ripeness is the trait a low-cost visible sensor genuinely recovers. On a benchmark that pools several fruit, a single accuracy conflates two questions, which fruit is this and what state is it in, and species is the louder signal. Species is nearly perfectly decodable from the spectrum, so the shortcut is always available, and what differs between the traits is whether each forces a model to read the spectrum beyond it. A species-lookup baseline separates the two, and ripeness clears it by a wide margin in every fruit, holds up when the prediction is repeated inside a single species, and survives folds grouped by physical fruit. A grader that sorts on ripeness therefore rests on a real measurement rather than on recognising the fruit. We recommend the species-lookup baseline as a default control for any multi-species quality benchmark, because accuracy reported without it can read as quality prediction when it is closer to fruit recognition.

Firmness carries a narrower lesson about labels rather than about sensing. Under one firmness scheme shared across all five fruit, a model gains little over the species baseline, which invites the reading that the spectra cannot see firmness. That reading is wrong. When firmness is labelled on each fruit's own scale, whether by splitting a fruit's own range into thirds or by the thresholds the dataset defines for avocado and kiwi, the same spectra recover firmness with a clear margin, consistent with spectral imaging predicting firmness within a single fruit type (Lu, 2004). The weak pooled firmness result is manufactured by harmonising a continuous mechanical property into classes that happen to line up with species. A benchmark that pools species should either label each fruit on its own scale or report firmness skill above a species baseline, so that a labelling choice is not mistaken for a sensing limit.

The biology supports reading ripeness as the real trait. Once species is controlled, the bands that carry ripeness concentrate near chlorophyll-a absorption and the red edge, the part of the spectrum that changes as chlorophyll degrades and colour turns during ripening. These are visible-range features, which is why a low-cost visible sensor is enough to recover ripeness. The same analysis warns against over-reading wavelength importance on pooled data. The pooled ranking spreads across the visible range and does not match the within-species ranking, so a band called important on a mixed-species set may be separating fruit rather than quality. Importance earns a physiological reading only after a within-species control.

The low-cost sensing story is real but narrower than a wavelength-slice

experiment suggests. Three bands recover most of the ripeness signal, yet the choice of three well-separated visible bands matters more than their precise placement, so an importance-guided triple is not special. Moving from clean spectral slices to a modelled sensor with a finite passband, photon noise, and illumination variation costs a large part of the accuracy. A deployable claim needs a real sensor and real illumination, not slices from a laboratory cube. We frame this as the next step rather than a defect, because the visible-range target it points to is cheap and reachable.

The central result is structural rather than incidental. A pooled accuracy is inflated whenever species is decodable from the spectrum and the label tracks species, and we measure both, and the ordering holds across every fruit subset we test, so the inflation follows from properties of the data rather than from a chance alignment in one benchmark or one mix of fruit. The species-lookup baseline that exposes it needs only species labels and applies to any multi-species quality benchmark. Read this way, the contribution is a demonstration of an evaluation pitfall and a control for it, shown on the field's standard set, rather than a claim that a single accuracy figure transfers to new fruit. The choice of the visible range is deliberate for the same reason a grader would make it, since a visible sensor is the cheap and deployable option, and the question we answer is what that sensor can honestly measure.

This study has limits that shape the claim. It uses one dataset and one camera in the visible and short near-infrared range, so the magnitudes we report are specific to this sensor and this set of fruit, and cross-dataset replication is the clear next test, since spectral calibrations are known to shift across instruments and sample sets (Feudale et al., 2002). The ripeness labels are ordinal and rest partly on visual judgement. A visible-band model may recover part of the cue a human annotator used, so a fully independent ripeness reference, such as an instrumented maturity index, would separate the physiological signal from the annotation process more cleanly than we can here. The firmness reading is specific to a sensor that cannot see the water and tissue bands past 1200 nm where mechanical properties are most legible. Whether a true near-infrared sensor recovers firmness that the visible range misses is an open and worthwhile question, and the species-lookup control we introduce is the tool to answer it honestly.

5. Conclusion

Pooling several fruit into one benchmark lets a model score quality by first recognising the fruit, and a single accuracy hides this. A baseline that predicts from species alone separates real spectral skill from fruit recognition. Read this way, visible spectra recover ripeness across five fruit through the pigments of ripening, and a low-cost visible sensor is a trustworthy and inexpensive route to ripeness grading. Firmness is recoverable too once it is labelled on each fruit's own scale, so the weak firmness result a pooled benchmark reports reflects cross-species labelling rather than the sensor. Any multi-species quality benchmark should report skill above a species baseline rather than accuracy alone.

Data and code availability

The DeepHS Fruit dataset (Varga et al., 2021) is public. The analysis code and the scripts that regenerate every table and figure will be made available upon publication.

CRediT authorship contribution statement

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

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare no competing financial interests or personal relationships that could have influenced this work.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used a large language model to assist with drafting and code development. After using this tool the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A. Model zoo and search spaces

Table A.5 lists the twenty classifiers and the hyperparameter grid tuned for each one. The grids are small on purpose. The pooled set holds roughly six hundred samples, so a wide grid would report tuning noise rather than generalisation. Each model sits behind a shared preprocessing prefix that the inner loop tunes jointly with the model, so the reported score reflects a preprocessing choice made without ever seeing the fold it is scored on.

Table A.5: The twenty classifiers and their hyperparameter search spaces, selected jointly with the shared preprocessing prefix inside the inner cross-validation loop. Grids are deliberately compact because the pooled set holds roughly six hundred samples, so a wide grid buys no generalisation evidence. The estimator behind each name is the corresponding scikit-learn, XGBoost, or LightGBM class in the released code. Every model also shares a preprocessing prefix tuned in the same loop, namely a standard scaler or passthrough, a class-balancing step (none, SMOTE, or random undersampling), and principal component analysis (none or 95% retained variance).

Model	Hyperparameter search space
Extra Trees	n_estimators=300; max_depth={None, 20}; min_samples_leaf={1, 2}
Random Forest	n_estimators=300; max_depth={None, 20}; min_samples_leaf={1, 2}
XGBoost	n_estimators=200; max_depth={4, 6}; learning_rate=0.1
LightGBM	n_estimators=200; num_leaves={31, 63}; learning_rate=0.1
HistGradientBoosting	max_depth={None, 6}; learning_rate=0.1; max_iter=200
SVC (RBF)	C={1.0, 10.0}; gamma=scale

continued on next page

Model	Hyperparameter search space
SVC (linear)	C={0.1, 1.0, 10.0}
k-NN	n_neighbors={5, 9}; weights={uniform, distance}
Logistic Regression	C={0.1, 1.0, 10.0}; penalty=l2
PLS-DA	n_components={3, 5, 10, 20, 40}
AdaBoost	n_estimators={100, 200}; learning_rate={0.5, 1.0}
Bernoulli NB	alpha={0.1, 1.0, 10.0}
Decision Tree	max_depth={None, 10, 20}; min_samples_leaf={1, 5}
Gaussian NB	var_smoothing={1e-09, 1e-07, 1e-05}
Gradient Boosting	n_estimators={100, 200}; max_depth={3, 5}; learning_rate=0.1
LDA	solver=svd
MLP	hidden_layer_sizes={(64,), (128,), (64, 32)}; alpha={0.0001, 0.001}
Passive-Aggressive	C={0.1, 1.0, 10.0}
Ridge	alpha={0.1, 1.0, 10.0}
SGD	loss={hinge, log_loss}; alpha={0.0001, 0.001}

Appendix B. Full model ranking

Table B.6 extends the ten-model view in Table 1 to the full field of twenty, for both tasks and with balanced accuracy (Equation 2) alongside macro-F_1. The task gap and the tree ensemble ordering hold across the whole field, and the confidence intervals confirm that the leaders on each task are not separated from one another. The point of Section 3 is that this ranking, read on its own, still overstates firmness until the species-lookup baseline is subtracted.

Table B.6: All twenty models under pooled nested cross-validation, ordered by ripeness macro-F_1. Each score is the mean over the ten outer folds with a 95% confidence interval, alongside balanced accuracy. Preprocessing is selected inside the inner folds, so the outer folds score only.

	Ripeness		Firmness	
Model	F_1 [95% CI]	bal. acc	F_1 [95% CI]	bal. acc
Extra Trees	0.666 [0.63, 0.70]	0.665	0.832 [0.79, 0.87]	0.833
Random Forest	0.653 [0.63, 0.68]	0.655	0.824 [0.78, 0.86]	0.823
LightGBM	0.652 [0.63, 0.68]	0.649	0.813 [0.77, 0.86]	0.816
HistGradientBoosting	0.629 [0.60, 0.65]	0.627	0.806 [0.76, 0.84]	0.808
Gradient Boosting	0.628 [0.60, 0.65]	0.629	0.796 [0.75, 0.84]	0.800
SVC (RBF)	0.626 [0.60, 0.65]	0.653	0.824 [0.79, 0.86]	0.815
XGBoost	0.625 [0.61, 0.64]	0.627	0.796 [0.74, 0.84]	0.798
k-NN	0.612 [0.56, 0.66]	0.607	0.813 [0.76, 0.86]	0.815
LDA	0.610 [0.58, 0.64]	0.619	0.792 [0.76, 0.82]	0.790
Ridge	0.599 [0.56, 0.64]	0.615	0.805 [0.77, 0.84]	0.803
PLS-DA	0.594 [0.57, 0.62]	0.603	0.828 [0.80, 0.86]	0.823
MLP	0.588 [0.54, 0.63]	0.612	0.720 [0.69, 0.75]	0.731
SVC (linear)	0.580 [0.54, 0.62]	0.600	0.788 [0.76, 0.82]	0.790
Logistic Regression	0.570 [0.53, 0.60]	0.581	0.779 [0.75, 0.81]	0.779
AdaBoost	0.545 [0.53, 0.57]	0.560	0.742 [0.69, 0.79]	0.744
Decision Tree	0.534 [0.49, 0.59]	0.543	0.729 [0.70, 0.76]	0.731
SGD	0.478 [0.43, 0.53]	0.500	0.704 [0.66, 0.75]	0.705
Gaussian NB	0.432 [0.40, 0.47]	0.469	0.649 [0.61, 0.69]	0.670
Bernoulli NB	0.432 [0.41, 0.46]	0.469	0.654 [0.61, 0.70]	0.674
Passive-Aggressive	0.396 [0.35, 0.45]	0.407	0.590 [0.51, 0.67]	0.613

Appendix C. Reproducibility

Every table and figure in this paper regenerates from the released scripts against the frozen result files. The evaluation uses a fixed seed of 42 throughout, an outer loop of ten stratified folds that score only, and an inner loop of five folds that runs a grid search over the shared preprocessing prefix and the model grid, selecting on macro-F_1. Parallelism lives at the grid-search level, so each estimator runs single-threaded to avoid nested oversubscription. Table C.7 records the software environment.

Table C.7: Software environment used to compute every result and to regenerate the tables and figures.

Component	Version
Python	3.11.15
scikit-learn	1.9.0
NumPy	2.4.6
SciPy	1.17.1
pandas	3.0.3
imbalanced-learn	0.14.2
XGBoost	3.2.0
LightGBM	4.6.0

References

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. Journal of Chemometrics 17, 166–173. doi:10.1002/cem.785.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy 43, 772–777. doi:10.1366/0003702894202201.

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.

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

Cawley, G.C., Talbot, N.L.C., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 11, 2079–2107.

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.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297. doi:10.1007/BF00994018.

Ding, G., Jin, K., Chen, X., Li, A., Guo, Z., Zeng, Y., 2024. Non-destructive prediction of ready-to-eat kiwifruit firmness based on Fourier transform near-infrared spectroscopy. Postharvest Biology and Technology 212, 112908. doi:10.1016/j.postharvbio.2024.112908.

Ezenarro, J., Schorn-García, D., 2025. How are chemometric models validated? a systematic review of linear regression models for NIRS data in food analysis. Journal of Chemometrics 39, e70036. doi:10.1002/cem.70036.

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.

Feudale, R.N., Woody, N.A., Tan, H., Myles, A.J., Brown, S.D., Ferré, J., 2002. Transfer of multivariate calibration models: a review. Chemometrics and Intelligent Laboratory Systems 64, 181–192. doi:10.1016/S0169-7439(02)00085-0.

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 20, 1–81.

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence 2, 665–673. doi:10.1038/s42256-020-00257-z.

Geladi, P., MacDougall, D., Martens, H., 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. Applied Spectroscopy 39, 491–500. doi:10.1366/0003702854248656.

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

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.

Király, P., Tóth, G., 2025. Being aware of data leakage and cross-validation scaling in chemometric model validation. Journal of Chemometrics 39, e70026. doi:10.1002/cem.70026.

Lee, J.E., Kim, M.J., Lee, B.Y., Hwan, L.J., Yang, H.E., Kim, M.S., Hwang, I.G., Jeong, C.S., Mo, C., 2025. Evaluating ripeness in post-harvest stored kiwifruit using VIS-NIR hyperspectral imaging. Postharvest Biology and Technology 225, 113496. doi:10.1016/j.postharvbio.2025.113496.

Li, H., Zhu, L., Li, N., Liu, Z., Wang, L., Chitrakar, B., Xu, D., Cui, Z., Tang, Y., Hu, L., Mo, H., 2024. NIR spectroscopy for quality assessment and shelf-life prediction of kiwifruit. Postharvest Biology and Technology 218, 113201. doi:10.1016/j.postharvbio.2024.113201.

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

Lu, R., 2004. Multispectral imaging for predicting firmness and soluble solids content of apple fruit. Postharvest Biology and Technology 31, 147–157. doi:10.1016/j.postharvbio.2003.08.006.

Lu, Y., Saeys, W., Kim, M., Peng, Y., Lu, R., 2020b. Hyperspectral imaging technology for quality and safety evaluation of horticultural products: A review and celebration of the past 20-year progress. Postharvest Biology and Technology 170, 111318. doi:10.1016/j.postharvbio.2020.111318.

Merzlyak, M.N., Gitelson, A.A., Chivkunova, O.B., Rakitin, V.Y., 1999. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. Physiologia Plantarum 106, 135–141. doi:10.1034/j.1399-3054.1999.106119.x.

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.

Nicolaï, B.M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K.I., Lammertyn, J., 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest Biology and Technology 46, 99–118. doi:10.1016/j.postharvbio.2007.06.024.

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.

Porat, R., Lichter, A., Terry, L.A., Harker, R., Buzby, J., 2018. Postharvest losses of fruit and vegetables during retail and in consumers' homes: Quantifications, causes, and means of prevention. Postharvest Biology and Technology 139, 135–149. doi:10.1016/j.postharvbio.2017.11.019.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 36, 1627–1639. doi:10.1021/ac60214a047.

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.

Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7, 91. doi:10.1186/1471-2105-7-91.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58, 109–130. doi:10.1016/S0169-7439(01)00155-1.

Zou, X., Zhao, J., Povey, M.J.W., Holmes, M., Mao, H., 2010. Variables selection methods in near-infrared spectroscopy. Analytica Chimica Acta 667, 14–32. doi:10.1016/j.aca.2010.03.048.