Meat Research
Vol 6, Issue 1: 1-11, Article 135, February, 2026
ISSN: 2790-1971
https://doi.org/10.55002/mr.6.1.135
http://www.bmsa.info/meatresearch

[1]Department of Animal Science, Bangladesh Agricultural University, Mymensingh-2202, Bangladesh

***Corresponding Author:**
MA Hashem
E-mail: hashem_as@bau.edu.bd

Keywords:
Cattle hide
Collagen thermal stability
Confound partitioning
Nested cross-validation
Sensory prediction
Machine learning

Article Info:
Received: December 15, 2025
Accepted: January 23, 2026
Published online: February 28, 2026

Research Article

Separating prediction from maturity in cooked cattle hide: confound-partitioned machine learning reveals a bounded thermal-lability axis

MM Rahman[1], A Ahmed[1], S Rahman[1], S Nusrat[1], A Nag[1], MA Hashem[1]*

Abstract

Small food-quality datasets can yield high machine-learning coefficients when a discrete biological factor separates samples, even if individual-level prediction is weak. Using cooked cattle hide as a collagen-dominated test system, we evaluated whether instrumental models predict trained-panel sensory quality beyond dentition class and whether predictive relationships persist across maturity domains. Neck hides from 24 male Sahiwal-crossbred cattle (two-tooth, n=12; four-tooth, n=12) was processed identically and characterized for pH, CIE L*a*b*, drip and cooking loss, Warner-Bratzler shear force, proximate composition and five sensory attributes. A seven-rung model ladder was evaluated by nested leave-one-animal-out cross-validation against intercept-only and dentition group-mean benchmarks. PC1 explained 78.70% of variance and the five core predictors were strongly collinear (VIF 11.15-43.03). Best cross-validated R^2 values were 0.980 for juiciness, 0.976 for texture and overall acceptability, 0.962 for color and 0.772 for flavor. For PLSR, fold-safe class centering reduced R^2 by only 5-27% (Φ=0.05-0.27), but every cross-class transfer had negative R^2. Exact retraining Shapley attribution across four physical blocks was nearly uniform (22.3-27.6%), precluding a unique instrumental driver. Cooking loss and shear force were correlated in younger hide (r=0.905) but not mature hide (r=-0.146); the correlations differed significantly (Fisher Z=3.49, P=0.00048). Thus, sensory prediction is substantial within the observed maturity structure but is not transportable across it, and the proposed thermal-lability axis is explicitly maturity-bounded.

Introduction

Objective prediction of eating quality is moving rapidly from laboratory assays toward spectroscopy, sensor fusion and machine-learning workflows. Recent Meat Science studies illustrate both the opportunity and the necessary restraint: electronic-tongue measurements have predicted trained-panel changes in vacuum-packed beef (Wilson et al., 2025), whereas a large REIMS study found that some tenderness classifiers could not outperform a no-information benchmark even with thousands of carcasses (Loomas et al., 2025). In 2026, REIMS signals were shown to track several biochemical tenderness markers, including soluble and insoluble collagen, but with markedly different coefficients across mechanisms (Loomas et al., 2026). These studies make a central point: useful prediction requires not only a high headline metric, but a clear benchmark, a defined calibration domain and a biologically credible interpretation. That requirement is especially important when n is small and the design contains discrete groups. Leakage through feature selection, repeated units or preprocessing can inflate apparent performance, with smaller datasets particularly vulnerable (Hossain et al., 2021 and 2023; Rosenblatt et al., 2024). Even confound correction can itself leak information when performed incorrectly (Hamdan et al., 2023). A second, distinct problem remains after leakage is controlled: a model can exploit between-group separation and still predict poorly within either group. In food studies this can occur whenever treatment, breed, storage state, authenticity class or physiological maturity shifts both predictors and outcomes. A pooled R^2 can therefore answer the wrong question.

Cooked cattle hide provides a stringent biological test because dermis is dominated by extracellular matrix, particularly type-I collagen, rather than by the myofibrillar and myoglobin systems that dominate muscle-meat interpretation. Evidence from bovine hide itself shows that hydration and cross-linking materially alter collagen thermal denaturation behavior (Schroepfer and Meyer, 2017), while edible bovine-hide collagen casings link swelling, solubility and shrinkage temperature with mechanical and sensory properties (Zając et al., 2021). Across beef muscles, collagen content and pyridinoline cross-links are positively associated with shear resistance, whereas collagen solubility is negatively associated with shear force; these relationships also vary with age and muscle (Li et al., 2022). Together, these findings justify a deliberately provisional thermal-lability framework: cooking loss, drip loss and shear force may partly reflect a common maturity-dependent response of the collagen network to a fixed heat load.

The present study therefore asks a stricter question than whether machine learning can fit sensory scores. We test whether instrumental prediction adds information beyond dentition class, whether that performance survives removal of the class offset, whether models transfer across maturity classes, and whether correlated predictors can be attributed without misleading single-variable rankings. We further examine an explicit boundary condition: if thermal water release and matrix mechanics are projections of one latent process, their correlation should persist within each maturity class. The resulting framework combines a group-mean benchmark, a fold-safe within-class diagnostic, cross-class transfer, dependence-aware block attribution and a target-reliability sensitivity check. This combination is intended as an auditable reporting workflow rather than as a new causal estimator.

Materials and methods

Animals, sampling and thermal processing

Neck hide was collected from 24 male Sahiwal-crossbred cattle slaughtered at a commercial facility in Islampur, Jamalpur, Bangladesh, with 12 animal-level experimental units in each dentition class. Dentition was classified by permanent incisor eruption: T1 comprised two-tooth cattle (one erupted pair; approximate age below two years) and T2 four-tooth cattle (two erupted pairs; approximate age two to four years). The individual animal was the unit of analysis throughout, and every outer cross-validation split held out one animal. The material originated from a broader quality-characterization trial; the present study addresses the distinct question of individual sensory prediction and maturity-confound partitioning rather than the group-contrast question itself. Hide was washed at 60-70 °C, dehaired, trimmed of subcutaneous tissue, frozen at -18 °C, thawed at 4 °C for 24 h and cooked in a water bath at 80 °C for 40 min. This intermediate sub-boiling treatment is above the principal denaturation/shrinkage range reported for hydrated collagenous tissues and provides a fixed thermal challenge without assuming complete collagen solubilization (Tornberg, 2005; Schroepfer and Meyer, 2017). The processing condition was therefore used to expose maturity-associated material differences, not as evidence that specific covalent cross-links were ruptured.

Physicochemical and compositional measurements

pH was measured by direct insertion of a calibrated probe. Surface color was recorded as CIE L*, a* and b* with a Konica Minolta CR-400 colorimeter calibrated against a standard white plate, using three readings per sample. Drip loss was determined gravimetrically over thawing following the general approach of Honikel (1998), and cooking loss was calculated from pre- and post-cooking mass. Warner-Bratzler shear force was measured on 1 × 1 × 2.5 cm strips and expressed in newtons. Proximate composition followed AOAC procedures: dry matter by oven drying at 105 °C, crude protein by Kjeldahl (N × 6.25), ether extract by Soxhlet extraction and ash by incineration at 550 °C (AOAC, 1995, 16th ed.). Carbohydrate was obtained by difference.

Sensory evaluation

Five assessors, screened and trained in accordance with ISO 8586 (2023) and meat sensory guidance, evaluated coded samples in individual booths under controlled conditions consistent with ISO 8589 (2007). Color, texture, flavors, juiciness and overall acceptability were scored on a five-point quality scale (5=excellent; 1=poor). Samples were presented with three-digit random codes at a uniform serving temperature, with water supplied between samples. The mean of the five assessors was the animal-level response. The panel was trained rather than consumer-based; therefore, the scores are interpreted as product-quality ratings rather than population liking. Informed consent to participate in the trained sensory panel was obtained from each assessor.

Predictor architecture and model ladder

Five predictors were specified for the core regression ladder: pH and crude protein (matrix chemistry), cooking loss and drip loss (thermal water release), and shear force (matrix mechanics). The three CIE color coordinates were excluded from this core ladder to preserve the pre-specified complexity budget and were introduced only in the separate eight-variable block-attribution analysis. No automated feature screening, polynomial expansion or interaction search was performed.

The formal model ladder was intentionally parsimonious: rung 0, an intercept-only null; rung 1, a dentition group-mean predictor fitted from the outer-training animals; rung 2, ordinary least squares using shear force alone; rung 3, multiple linear regression; rung 4, PLSR; rung 5, Elastic Net; and rung 6, random forest. This sequence tests whether prediction exists, whether instrumental information improves on dentition alone, and whether added nonlinear flexibility improves generalization. Hyperparameter tuning was confined to the inner cross-validation loop (Table 1).

Table 1: Hyperparameter search space. Search was confined to inner cross-validation; the outer held-out animal was never used for selection

Model	Hyperparameter	Search grid	Selection criterion
Partial least squares	Latent variables	1, 2, 3	Inner 5-fold RMSE
Elastic Net	α	10^{-3} to 10^{1}, 12 points, log-spaced	Inner 5-fold RMSE
Elastic Net	L1 ratio	0.1, 0.5, 0.9	Inner 5-fold RMSE
Random forest	Trees	300 (fixed)	—
Random forest	Maximum depth	2, 3, unlimited	Inner 5-fold RMSE
Random forest	Maximum features	0.5, 1.0	Inner 5-fold RMSE
Random forest	Minimum samples per leaf	1, 2, 3	Inner 5-fold RMSE

Nested validation, performance and confound partitioning

Performance was estimated by nested leave-one-animal-out (LOAO) cross-validation. For each of 24 outer folds, one animal was held out and the remaining 23 formed the training set. PLSR, Elastic Net and random-forest hyperparameters were selected by five-fold cross-validation inside those 23 animals, the selected pipeline was refitted on all 23, and the held-out animal was predicted once. Standardization was fitted inside the training pipeline. Metrics were then computed from the 24 strictly out-of-fold predictions. This architecture separates model selection from performance estimation, a requirement that is especially important in limited datasets (Cawley and Talbot, 2010; Varma and Simon, 2006; Vabalas et al., 2019).

$$R^2 = 1 - \Sigma (y_i - \hat{y}_i)^2 / \Sigma (y_i - \bar{y})^2; \quad RMSE = [\Sigma (y_i - \hat{y}_i)^2 / n]^{1/2}$$

Two additional regimes tested the maturity-class dependence of the pooled result. First, predictors and responses were centered by dentition class using means estimated from the outer-training data only; the within-class R^2 used the within-class total sum of squares, $\Sigma_k \Sigma_{i \in T_k} (y_{ki} - \bar{y}_k)^2$. Second, each class was used as a training domain and the other as a test domain. We define $\Phi = 1 - R^2_{within} / R^2_{pooled}$ as the relative loss of R^2 after removal of the class offset. When both R^2 values are non-negative and $R^2_{within} \leq R^2_{pooled}$, Φ lies between 0 and 1. It is not a causal variance decomposition: it can be negative if centering improves performance and can exceed 1 if within-class R^2 is negative. Φ is therefore interpreted only as a model- and response-specific diagnostic and always alongside the group-mean floor and cross-class transfer.

$\Phi = 1 - R^2_{within} / R^2_{pooled}$

Significance, uncertainty, dependence-aware attribution and reliability

For the PLSR pipeline, the response was permuted 1000 times and the complete nested LOAO procedure, including all fold-specific preprocessing and tuning, was repeated to form an empirical null distribution; P=(1+b)/(1+B), where B=1000 and b is the number of permuted R^2 values at least as large as observed. Bias-corrected accelerated bootstrap intervals were calculated from 2000 resamples of the stored out-of-fold prediction pairs, and learning curves were constructed by repeated subsampling to 8, 12, 16, 20 and 24 animals. The bootstrap intervals summarize uncertainty in the observed out-of-fold prediction pairs and do not substitute for external-validation uncertainty. Negative mean R^2 under permutation is expected for out-of-fold models trained on noise and is not, by itself, evidence against leakage; leakage control follows from fold isolation.

Severe predictor dependence required a block-level rather than single-feature interpretation. Exact retraining Shapley values were calculated over four pre-defined blocks: thermal water release (cooking and drip loss), matrix mechanics (shear force), matrix chemistry (pH and crude protein), and optical properties (L*, a*, b*). For all 2^4=16 coalitions, a model was refitted and cross-validated; $v(S)=\max[0, R^2(S)]$ and $v(\emptyset)=0$. Retraining avoids the synthetic feature combinations created by independence-based perturbation methods, but it does not make attribution causal or eliminate ambiguity among correlated proxy blocks (Aas et al., 2021; Hooker et al., 2021).

$\varphi_j = \Sigma S{\subseteq}N\backslash\{j\}\ |S|!\ (p-|S|-1)!\ /p! \times [v(S{\cup}\{j\})-v(S)],\ p=4$

Finally, assessor-level ballots were unavailable, so ICC (2, k) for the panel mean could not be estimated. We therefore report a classical measurement-error sensitivity analysis. Under the assumption that panel reliability is the fraction of observed-score variance attributable to the latent true score, R^2observed/reliability >1 indicates incompatibility between the observed model coefficient and that assumed reliability. Such a value is a diagnostic signal—not automatic proof of overfitting—because sampling uncertainty, model optimism, reliability misspecification or departures from the classical error model can also contribute.

Univariate statistics and software

Group means were compared with Welch's t-test when Levene's test indicated unequal variances and Student's t-test otherwise; non-parametric checks were performed and P values were controlled across traits by the Benjamini-Hochberg procedure. Effect sizes are reported as Cohen's d. Pearson correlations were calculated pooled and separately within T1 and T2. Independent class-specific correlations were compared by Fisher's r-to-z transformation. Analyses were performed in Python 3.11 using scikit-learn 1.5, SciPy 1.14, NumPy 1.26 and pandas 2.2, with random seed 42.

Results and discussion

Maturity creates a two-cluster quality landscape

Dentition class separated almost every measured property (Table 2). T2 hide had higher pH, L*, drip loss, cooking loss and shear force, lower crude protein and higher ether extract, and lower scores for every sensory attribute. The largest standardized contrasts were shear force (d=-12.10), ether extract (d=-10.16), overall acceptability (d=6.69) and pH (d=-6.08). Redness and yellowness were the two exceptions. Such extreme separation is biologically informative but statistically hazardous: any flexible predictor can use the class structure as a shortcut unless performance is benchmarked against a predictor that already knows dentition.

The physical pattern is compatible with age-related collagen stabilization, but the present data do not identify specific cross-link species. Direct bovine-hide studies demonstrate that both cross-linking and hydration alter thermal denaturation behavior (Schroepfer and Meyer, 2017), and edible hide-collagen casings show coupled changes in swelling, solubility, shrinkage temperature and mechanical response (Zając et al., 2021). In meat, meta-analysis further shows that collagen content, solubility and pyridinoline associations with tenderness depend on age and muscle (Li et al., 2022). The current results therefore support a collagen-centered hypothesis without establishing a specific stereochemical pathway. The measured pH values (6.768 and 7.646) should likewise not be interpreted through the classic post-rigor muscle DFD mechanism: hide dermis is not a myofiber-dominated system, and controlled pH manipulation was not performed. pH is treated here as a maturity-associated matrix marker rather than a demonstrated cause.

Table 2: Effect of dentition class on physicochemical, compositional and trained-panel sensory attributes of cooked cattle hide (mean ± SEM; n=12 animals per class)

Attribute	T_1 (2-tooth)	T_2 (4-tooth)	Levene P	Test	P	Cohen's d	CV T_1 (%)	CV T_2 (%)
Physicochemical								
pH	6.768 ± 0.013	7.646 ± 0.057	0.003	Welch	<0.001	−6.08	0.68	2.60
L*	54.006 ± 1.026	62.399 ± 1.477	0.498	Student	<0.001	−1.91	6.58	8.20
a*	2.558 ± 0.191	2.068 ± 0.266	0.681	Student	0.149	0.61	25.82	44.58
b*	13.033 ± 0.176	11.936 ± 0.716	0.011	Welch	0.162	0.61	4.67	20.78
Drip loss (%)	8.354 ± 0.155	12.162 ± 0.562	0.013	Welch	<0.001	−2.67	6.44	16.02

Attribute	T_1 (2-tooth)	T_2 (4-tooth)	Levene P	Test	P	Cohen's d	CV T_1 (%)	CV T_2 (%)
Cooking loss (%)	4.589 ± 0.221	15.992 ± 1.150	0.002	Welch	<0.001	−3.98	16.69	24.91
Shear force (N)	154.63 ± 2.31	258.48 ± 2.63	0.696	Student	<0.001	−12.10	5.18	3.52
Proximate composition								
Dry matter (%)	36.192 ± 0.176	34.184 ± 0.122	0.173	Student	<0.001	3.83	1.69	1.23
Moisture (%)	63.808 ± 0.176	65.816 ± 0.122	0.173	Student	<0.001	−3.83	0.96	0.64
Crude protein (%)	33.465 ± 0.299	29.528 ± 0.144	0.036	Welch	<0.001	4.85	3.09	1.69
Ether extract (%)	1.565 ± 0.039	2.749 ± 0.027	0.145	Student	<0.001	−10.16	8.71	3.38
Ash (%)	0.686 ± 0.039	1.158 ± 0.047	0.654	Student	<0.001	−3.17	19.56	14.08
Trained-panel sensory								
Color	4.133 ± 0.081	2.608 ± 0.122	0.200	Student	<0.001	4.25	6.79	16.22
Texture	4.275 ± 0.091	2.542 ± 0.087	0.712	Student	<0.001	5.62	7.41	11.80
Flavor	3.517 ± 0.074	2.775 ± 0.097	0.353	Student	<0.001	2.49	7.26	12.11
Juiciness	4.467 ± 0.063	2.900 ± 0.170	0.008	Welch	<0.001	3.53	4.90	20.27
Overall acceptability	4.042 ± 0.051	2.683 ± 0.065	0.257	Student	<0.001	6.69	4.41	8.38

Note: Welch's correction was used where Levene's test indicated heterogeneity. Sensory scores: 5=excellent to 1=poor. CV, coefficient of variation. Moisture+crude protein+ether extract+ash equals 99.54% in T1 and 99.25% in T2; the small remainder corresponds to carbohydrate-by-difference and rounding.

A dominant but not automatically mechanistic latent axis

PCA of eleven instrumental and sensory variables yielded PC1=78.70% and PC2=12.47% of variance (Table 3; Fig. 1). Cooking loss (+0.331), pH (+0.330), shear force (+0.322) and drip loss (+0.313) opposed overall acceptability (-0.339), juiciness (-0.330), color (-0.326), texture (-0.315) and flavor (-0.291) on PC1. PC2 was driven mainly by L* (+0.621) and a* (+0.628). In dermis, these optical coordinates should be interpreted primarily as surface appearance and light-scattering responses to matrix structure and processing, not as myoglobin-redox markers. The PCA establishes strong low dimensionality; it does not prove that a single biochemical process caused 78.70% of the variance, because dentition itself can generate a dominant axis.

Table 3: PCA eigen structure and first two loading vectors for the eleven-variable analysis

Component	Eigenvalue	Variance (%)	Cumulative (%)	Variable	PC1 loading	PC2 loading
PC1	8.657	78.70	78.70	Cooking loss	+0.331	−0.128
PC2	1.371	12.47	91.17	pH	+0.330	−0.013
PC3	0.616	5.60	96.77	Shear force	+0.322	+0.231
PC4	0.200	1.81	98.58	Drip loss	+0.313	−0.219
PC5–PC11	0.156	1.42	100.00	L*	+0.206	+0.621
				a*	−0.155	+0.628
				Flavor	−0.291	+0.028
				Texture	−0.315	−0.266
				Color	−0.326	+0.117
				Juiciness	−0.330	+0.122
				Overall acceptability	−0.339	−0.037

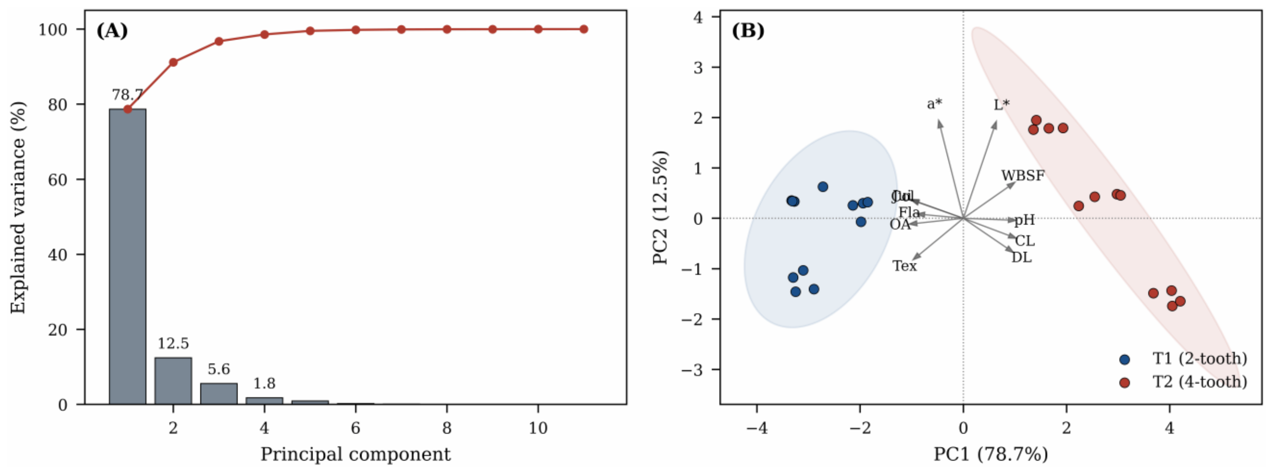

Figure 1: Latent structure of cooked cattle-hide quality. (A) Scree plot. (B) PC1-PC2 biplot with dentition-class ellipses. Thermal-loss/mechanical variables oppose sensory quality on PC1; optical variables contribute strongly to PC2. DL, drip loss; CL, cooking loss; WBSF, Warner-Bratzler shear force; Col, color; Tex, texture; Fla, flavor; Jui, juiciness; OA, overall acceptability.

Predictor collinearity changes what can be claimed

Variance inflation factors for the five core predictors were 43.03 for cooking loss, 35.46 for drip loss, 28.32 for shear force, 17.74 for pH and 11.15 for crude protein. This is not merely a regression nuisance. It means that each variable carries substantial information about the others, so a conventional importance ranking can assign apparently precise credit to measurements that are partially interchangeable. The appropriate inferential target is therefore the stability of physically coherent blocks, not a causal ranking of individual predictors.

The thermal-lability interpretation has a statistically testable boundary

Several pooled correlations remained strong within both classes (Table 4; Fig. 2): shear force versus texture (T1 r=-0.921; T2 r=-0.947), cooking loss versus juiciness (-0.939; -0.988), drip loss versus juiciness (-0.870; -0.875), pH versus color (-0.860; -0.882) and cooking loss versus overall acceptability (-0.806; -0.892). These repeated within-class relationships are stronger mechanistic candidates than pooled correlations alone.

Two relationships failed specifically in T2. Texture no longer tracked overall acceptability (r=-0.164, P=0.611), and—more important for the thermal-lability hypothesis—cooking loss no longer tracked shear force (r=-0.146, P=0.650), despite a strong T1 relationship (r=+0.905, P=0.0001). Fisher's comparison of the independent correlations gave Z=3.49 and two-sided P=0.00048. The difference is therefore not established merely by comparing two P values; the correlations themselves are statistically different. This result prevents the paper from claiming a universal one-axis mechanism. Mature hide may be mechanically consistent while water release remains heterogeneous: cooking-loss CV increased from 16.69% to 24.91%, whereas shear-force CV decreased from 5.18% to 3.52%. Direct hydroxyproline solubility, cross-link chemistry and water-mobility measurements are needed to resolve the molecular basis.

Table 4. Pooled and within-dentition Pearson correlations. P values are shown for class-specific coefficients

Relationship	**Pooled r**	**T_1 r (P)**	**T_2 r (P)**	**Interpretation**
Shear force × texture	−0.982	−0.921 (<0.001)	−0.947 (<0.001)	Holds within both classes
Cooking loss × juiciness	−0.993	−0.939 (<0.001)	−0.988 (<0.001)	Holds within both classes
Drip loss × juiciness	−0.956	−0.870 (0.0002)	−0.875 (0.0002)	Holds within both classes
pH × color	−0.971	−0.860 (0.0003)	−0.882 (0.0001)	Holds within both classes
Cooking loss × overall acceptability	−0.959	−0.806 (0.0015)	−0.892 (0.0001)	Holds within both classes
Texture × overall acceptability	+0.931	+0.729 (0.0071)	−0.164 (0.611)	Fails within T2
Cooking loss × shear force	+0.890	+0.905 (0.0001)	−0.146 (0.650)	Fails within T2; violates declared expectation

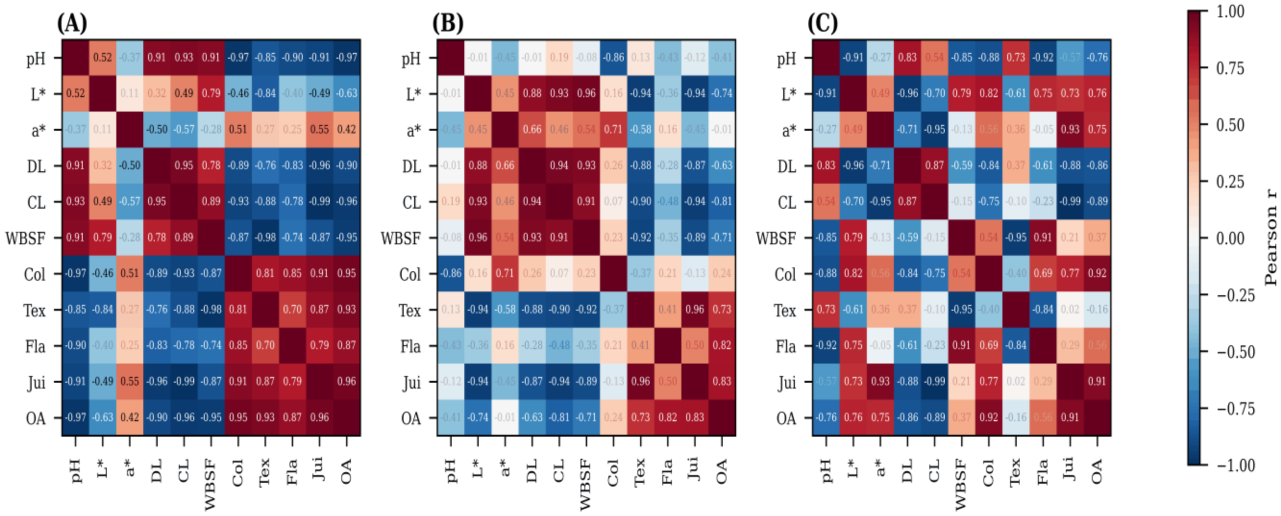

Figure 2. Pearson correlation structure for (A) all animals (n=24), (B) T_1 only (n=12) and (C) T_2 only (n=12). The common scale exposes relationships that persist after the maturity-class contrast is removed and relationships that do not.

High pooled R^2 is meaningful only relative to the dentition floor

The seven-rung ladder separates three questions: does any prediction exist, does instrumental information beat the freely observable maturity class, and does add model complexity help? Table 5 and Fig. 3 show that the intercept-only LOAO null gives R^2=-0.089 by construction, whereas the group-mean floor is already high for texture (0.876), color (0.799) and overall acceptability (0.910). A shear-force-only model beat that floor only for texture. In contrast, multivariable linear or regularized models added substantial information for juiciness (+0.250 at best) and flavor (+0.216), but much less for overall acceptability (+0.066). The practical value of an instrumental model is therefore attribute-specific, even when headline R^2 values are all high.

Regularized and linear models were best or essentially tied across responses. Best R^2 was 0.980 for juiciness (Elastic Net), 0.976 for texture (MLR/Elastic Net) and overall acceptability (PLSR), 0.962 for color (MLR), and 0.772 for flavor (PLSR). With n=24 and a strongly one-dimensional predictor space, this is the expected bias-variance trade-off: added nonlinear flexibility is not rewarded. The observation aligns with recent food-sensing work in which model ranking depended on the target rather than on algorithm prestige, and it contrasts usefully with very large REIMS datasets where even 13 algorithms could fail to beat a baseline for a particular tenderness categorization (Loomas et al., 2025).

Table 5: Nested LOAO model ladder for trained-panel sensory attributes. ΔR^2 is relative to the dentition group-mean floor

Attribute	**Model**	**R^2 [95% CI]**	**RMSE**	**ΔR^2 vs floor**
Texture	Intercept-only null	−0.089	0.955	—
	Group-mean floor	0.876 [0.830, 0.910]	0.322	0.000
	OLS, shear force only	0.957 [0.931, 0.972]	0.190	+0.081
	Multiple linear regression	0.976 [0.957, 0.988]	0.142	+0.100
	Partial least squares	0.969 [0.942, 0.984]	0.161	+0.093
	Elastic Net	0.976 [0.956, 0.987]	0.143	+0.099
	Random forest	0.970 [0.948, 0.984]	0.159	+0.093
Juiciness	Group-mean floor	0.730 [0.494, 0.836]	0.463	0.000
	OLS, shear force only	0.721 [0.522, 0.853]	0.470	−0.009
	Multiple linear regression	0.978 [0.955, 0.989]	0.134	+0.248
	Partial least squares	0.978 [0.957, 0.989]	0.133	+0.248
	Elastic Net	0.980 [0.963, 0.988]	0.125	+0.250
	Random forest	0.972 [0.946, 0.987]	0.149	+0.242
Flavor	Group-mean floor	0.557 [0.238, 0.716]	0.312	0.000
	OLS, shear force only	0.466 [0.078, 0.659]	0.342	−0.091
	Multiple linear regression	0.769 [0.508, 0.888]	0.225	+0.213
	Partial least squares	0.772 [0.578, 0.879]	0.223	+0.216

Attribute	Model	R² [95% CI]	RMSE	ΔR² vs floor
	Elastic Net	0.750 [0.570, 0.865]	0.234	+0.194
	Random forest	0.766 [0.503, 0.868]	0.226	+0.209
Color	Group-mean floor	0.799 [0.666, 0.860]	0.375	0.000
	OLS, shear force only	0.720 [0.584, 0.802]	0.442	−0.079
	Multiple linear regression	0.962 [0.927, 0.983]	0.163	+0.163
	Partial least squares	0.949 [0.905, 0.974]	0.190	+0.150
	Elastic Net	0.957 [0.920, 0.978]	0.172	+0.158
	Random forest	0.931 [0.876, 0.962]	0.220	+0.132
Overall acceptability	Group-mean floor	0.910 [0.863, 0.936]	0.212	0.000
	OLS, shear force only	0.892 [0.826, 0.934]	0.232	−0.018
	Multiple linear regression	0.970 [0.941, 0.983]	0.123	+0.060
	Partial least squares	0.976 [0.959, 0.986]	0.110	+0.066
	Elastic Net	0.971 [0.951, 0.983]	0.120	+0.061
	Random forest	0.974 [0.961, 0.984]	0.114	+0.064

For LOAO, the intercept-only R² is −0.089 for every response by construction and is shown once to avoid repetition.

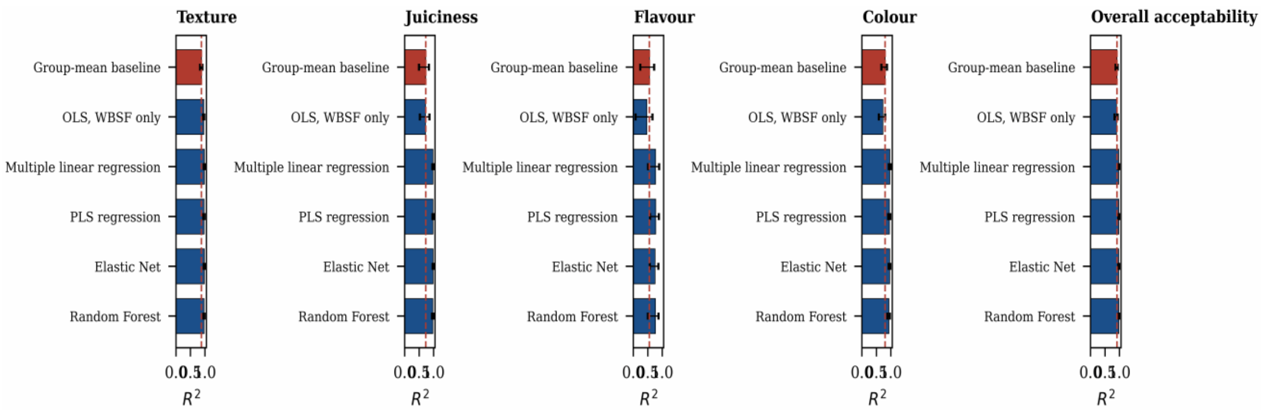

Figure 3: Cross-validated R² across the complete model ladder. The group-mean benchmark makes visible whether an instrumental model adds predictive value beyond dentition class alone.

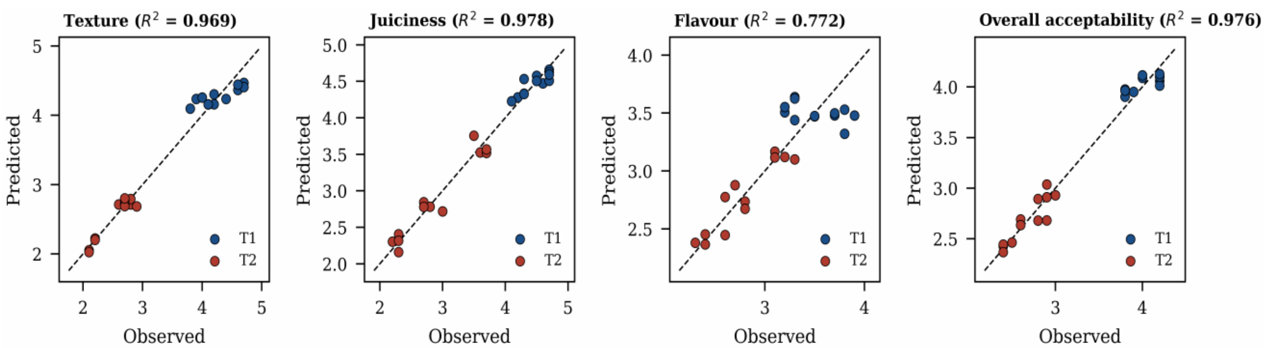

Figure 4: Out-of-fold PLSR predictions versus observed trained-panel scores for four key sensory attributes. Points are shown by dentition class and against the identity line. Color is omitted from the parity panel for visual economy; its metrics are fully reported in Table 5.

Confound partitioning distinguishes within-class skill from domain transfer

Class-centered evaluation changes the interpretation of the headline coefficients without nullifying them. For PLSR, Φ was 0.05 for juiciness, 0.12 for texture, 0.18 for color, 0.21 for overall acceptability and 0.27 for flavor; random forest reached 0.35 for flavor (Table 6; Fig. 5). Thus, much of the pooled R² survives removal of the class offset, especially for juiciness. This argues against a purely two-cluster explanation but does not make the remaining association causal: Φ is a relative performance-loss diagnostic calculated in the same small cohort.

The more severe result is cross-class transfer. Every T1→T2 and T2→T1 coefficient was negative, ranging from -0.02 to -46.27. A model can therefore interpolate within the calibrated maturity structure yet extrapolate disastrously outside it. This is a critical distinction for deployment. The data support dentition-stratified calibration, not a universal abattoir model. In that sense, failed transfer is not an embarrassing secondary result; it defines the model's domain of validity.

Table 6: Confound-partitioning diagnostic and cross-class transfer. Φ=1-R²within/R²pooled

Attributes	Model	Pooled R^2	Within-class R^2	Φ	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$
Juiciness	PLSR	0.978	0.930	0.05	−4.98	−0.02
	Random forest	0.972	0.926	0.05	−5.52	−21.88
Texture	PLSR	0.969	0.855	0.12	−31.44	−16.18
	Random forest	0.970	0.825	0.15	−27.31	−39.80
Color	PLSR	0.949	0.778	0.18	−5.58	−0.57
	Random forest	0.931	0.767	0.18	−13.51	−29.00
Overall acceptability	PLSR	0.976	0.776	0.21	−36.07	−9.57
	Random forest	0.974	0.712	0.27	−28.21	−46.27
Flavor	PLSR	0.772	0.567	0.27	−34.93	−7.14
	Random forest	0.766	0.500	0.35	−4.32	−4.15

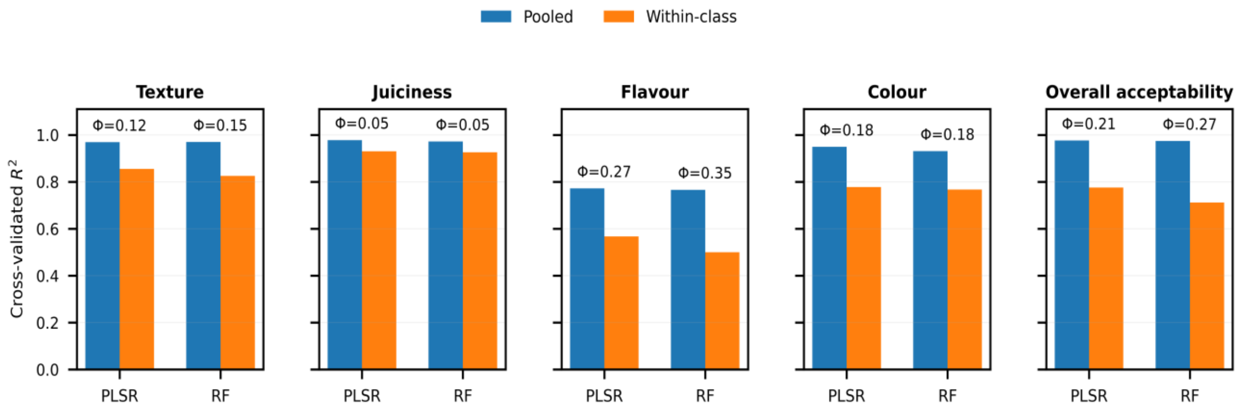

Figure 5: Confound-partitioning diagnostic. Bars compare pooled and fold-safe within-class R² for PLSR and random forest. Φ is the relative loss of R² after class centering when both coefficients are non-negative. The shared legend is placed outside the panels to prevent overlap with class/model labels.

Dependence-aware attribution rejects a single dominant instrumental driver

Exact four-block retraining Shapley attribution was strikingly uniform (Table 7; Fig. 6). Across the five sensory responses, block shares ranged only from 22.3% to 27.6%, with a maximum within-response spread of 4.3 percentage points. This pattern is a positive result: under severe collinearity, no measurement family earns robust dominance. Thermal water release, mechanics, matrix chemistry and optical properties are alternative views of a tightly coupled material state. The analysis therefore prevents a common interpretability error—presenting the largest importance bar as a unique biological driver when correlated features are substitutable.

Flavor is the most revealing case. It had the lowest overall predictive performance and the largest Φ, consistent with a predictor set designed around mechanics, hydration and bulk composition rather than volatile chemistry. Ether extract differed strongly between classes, but volatile oxidation and Maillard products were not measured. The correct inference is therefore that the present model predicts flavor partly through correlated maturity signatures; it does not identify flavor chemistry.

Table 7: Exact retraining Shapley attribution across four physically coherent predictor blocks

Attributes	Total R^2	Thermal water release	Matrix mechanics	Matrix chemistry	Optical	Spread (pp)
Texture	0.974	25.6	26.1	24.4	23.9	2.2
Juiciness	0.970	27.6	24.9	23.3	24.3	4.3
Flavor	0.789	24.8	24.6	26.6	23.9	2.7
Color	0.935	25.6	25.1	26.1	23.2	2.9
Overall acceptability	0.977	26.6	26.0	25.1	22.3	4.3

Percentages are rounded to one decimal place and may sum to 100.0 ± 0.1%.

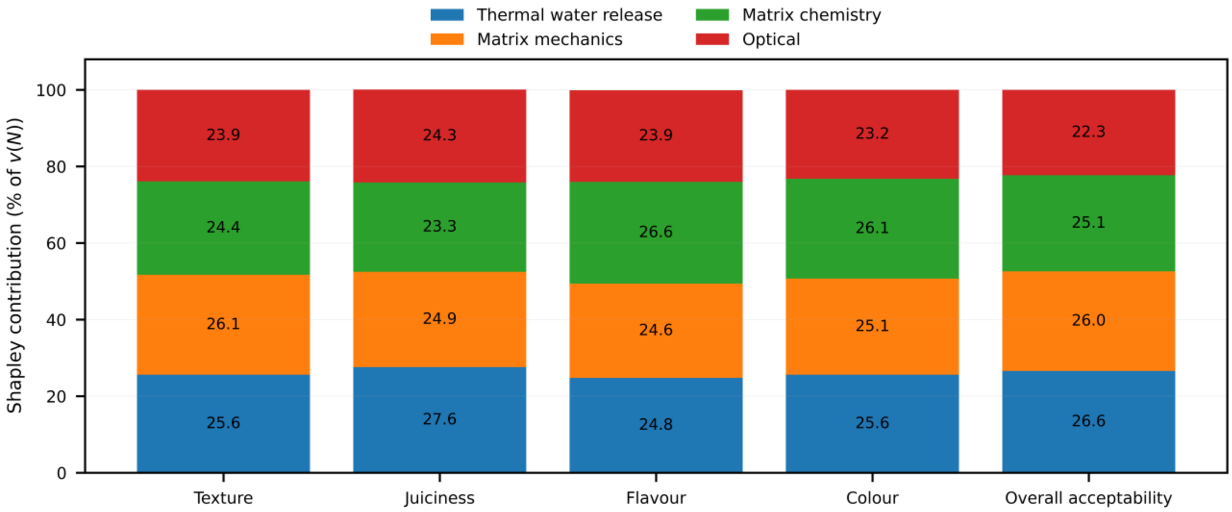

Figure 6. Exact block-level Shapley contributions expressed as percentages of v(N). Near-uniform shares under severe predictor dependence show why a unique single-variable driver cannot be defended from this dataset.

Statistical bounds keep the high coefficients provisional

PLSR performance was far outside the empirical permutation null for every response (Table 8; Fig. 7A); with 1000 permutations, P=0.001 is the minimum attainable corrected P value. The learning curves, however, were still rising at n=24 (Fig. 7B): texture increased from R^2=0.679 at n=8 to 0.969 at n=24, and overall acceptability from 0.911 to 0.976. The texture interval at n=8 was particularly wide (0.111-0.974). These coefficients should therefore be described as internally validated pilot estimates, not as final calibration accuracy.

Target reliability provides a separate interpretive bound (Table 9). Because assessor-level scores were not retained, no empirical ICC can be reported. Under a classical error model, a juiciness R^2 of 0.980 is incompatible with assumed panel reliability below approximately 0.98 in the narrow mathematical sense that R^2/reliability exceeds one. At assumed reliability 0.85, the ratio is 1.153. This does not establish overfitting by itself; it demonstrates why future sensory-ML studies should retain individual ballots, report an appropriate absolute-agreement ICC for the panel mean and propagate measurement uncertainty into model interpretation.

Table 8: Permutation testing for the PLSR pipeline (1000 response permutations; complete nested LOAO pipeline repeated for each permutation)

Attributes	Observed R^2	Null mean	Null 95th percentile	P
Texture	0.969	−0.336	0.032	0.001
Juiciness	0.978	−0.334	0.052	0.001
Flavor	0.772	−0.338	0.042	0.001
Color	0.949	−0.332	0.050	0.001
Overall acceptability	0.976	−0.339	0.027	0.001

Table 9: Reliability sensitivity analysis under a classical measurement-error model

Attributes	Best observed R^2	ICC 0.70	ICC 0.75	ICC 0.80	ICC 0.85	ICC 0.90	ICC 0.95
Juiciness	0.980	1.400	1.307	1.225	1.153	1.089	1.032
Texture	0.976	1.394	1.301	1.220	1.148	1.084	1.027
Overall acceptability	0.976	1.394	1.301	1.220	1.148	1.084	1.027
Color	0.962	1.374	1.283	1.203	1.132	1.069	1.013
Flavor	0.772	1.103	1.030	0.966	0.909	0.858	0.813

Entries are R^2observed divided by assumed reliability of the panel mean. Ratios >1 indicate incompatibility with that assumed reliability under the stated error model; they are diagnostic and are not, on their own, proof of overfitting.

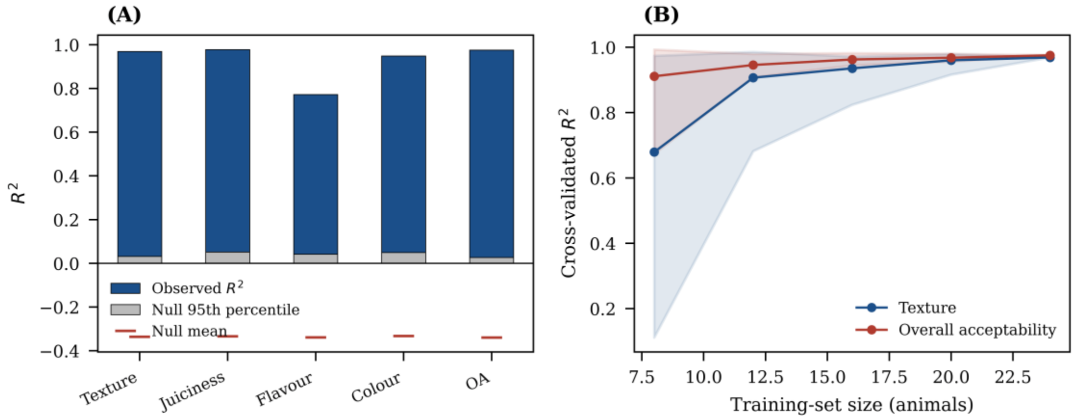

Figure 7: Statistical bounds on model interpretation. (A) Observed PLSR R^2 values compared with empirical permutation-null summaries. (B) Learning curves with bootstrap uncertainty; continued improvement at n=24 indicates that the calibration is not saturated. OA, overall acceptability.

General implications and falsifiable next tests

The analytical contribution is an auditable separation of questions that are often conflated: does a model beat a null predictor, does it beat a known-group benchmark, does predictive performance survive removal of the group offset, and does the relationship transfer beyond the calibration domain? A further check asks whether the sensory target is measured reliably enough to support the claimed coefficient. Together with exact block retraining under severe collinearity, these checks turn a strongly clustered dataset into a transparent test of predictive validity. The framework is operational rather than causal; each diagnostic can falsify an overly broad interpretation.

The biophysical result is deliberately narrower. A thermal-lability axis is plausible at the pooled level and within younger hide, but mature hide falsifies the simplest universal version because water release and mechanical resistance decouple. This is more informative than forcing a single collagen narrative. The next experiment is now sharply specified: quantify total and cooking-liquor hydroxyproline, collagen cross-link chemistry and DSC denaturation behaviour on the same biological units, and pair these with low-field NMR water populations. Recent Meat Science work demonstrates the value of linking predictive signals to independently measured biochemical tenderness mechanisms rather than treating prediction as mechanism (Loomas et al., 2026). A controlled pH-adjustment experiment should additionally test whether pH is a causal processing lever or a maturity marker.

Limitations

The study has 24 animals, two discrete maturity classes, one sex, one anatomical site and one thermal protocol. Learning curves remain unsaturated and there is no external farm-, day-, operator- or laboratory-level validation. Dentition is an imperfect proxy for chronological and biochemical maturity. Panel reliability cannot be estimated because individual assessor ballots were not retained. The collagen mechanism is inferred from physical measurements; hydroxyproline solubility, native cross-link species, DSC and NMR were not measured. The Φ statistic compares cross-validated coefficients referenced to different sums of squares and is a relative performance diagnostic rather than a formal partition of explained variance; it addresses the observed dentition grouping but cannot detect unknown confounders. Bootstrap intervals from stored out-of-fold predictions do not capture all model-development uncertainty. Finally, complete failure of cross-class transfer means that no industrial deployment claim is justified without calibration spanning the intended maturity distribution.

Conclusions

Cooked cattle hide produced very high internally cross-validated sensory predictions, but their scientific meaning became clear only after explicit benchmarking and partitioning. Multivariable linear/regularised models generally outperformed the nonlinear ensemble; instrumental measurements added substantially beyond dentition for juiciness and flavour but only modestly for overall acceptability. Within-class R^2 remained high for several responses, yet models failed completely when transferred between maturity classes. Severe collinearity distributed Shapley credit nearly uniformly across physical blocks, so no individual measurement can be defended as the unique driver. Most importantly, the strong T1 cooking-loss-shear-force relationship disappeared in T2, statistically bounding the proposed collagen thermal-lability axis. The resulting message is both methodological and biological: small-sample food ML should report a meaningful group benchmark, fold-safe within-group performance, domain transfer, dependence-aware attribution and target-reliability sensitivity; mechanistic claims should survive the same within-group tests as predictive claims.

Data availability

The anonymized animal-level dataset and analysis code supporting the findings are available from the corresponding author on reasonable request.

Ethics statement

Informed consent to participate in the trained sensory panel was obtained from each assessor.

Funding

This work was supported by the National Science and Technology (NST) Fellowship, Ministry of Science and Technology, Government of the People's Republic of Bangladesh, and by the Palli Karma-Sahayak Foundation (PKSF).

Acknowledgements

The authors thank the Department of Animal Science, Bangladesh Agricultural University, for laboratory facilities and the trained sensory panel for participation.

Declaration of competing interest

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

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used OpenAI ChatGPT to support manuscript organization, language refinement and consistency checking. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, 103502. https://doi.org/10.1016/j.artint.2021.103502

American Meat Science Association (AMSA). (1995). Research guidelines for cookery, sensory evaluation and instrumental tenderness measurements of fresh meat. National Livestock and Meat Board, Chicago, IL, USA.

AOAC. (1995). Official methods of analysis (16th ed.). Association of Official Analytical Chemists, Arlington, VA, USA.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/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.

Hamdan, S., Love, B. C., von Polier, G. G., Weis, S., Schwender, H., Eickhoff, S. B., & Patil, K. R. (2023). Confound-leakage: Confound removal in machine learning leads to leakage. GigaScience, 12, giad071. https://doi.org/10.1093/gigascience/giad071

Hashem, M. A., Ahmmed, A., & Hasan, M. M. (2025). Non-destructive discrimination of fresh, aged, and frozen-thawed beef using portable near-infrared spectroscopy combined with explainable artificial intelligence. LWT - Food Science and Technology, 236, 118659. https://doi.org/10.1016/j.lwt.2025.118659

Hashem, M. A., Akter, R., Ahmmed, A., Billah, M. M., & Rahman, M. M. (2024). Detection of adulteration of cattle and buffalo meat through NIRS and chemometric analysis. Meat Research, 4(3), Article 95. https://doi.org/10.55002/mr.4.3.95

Hashem, M. A., Hossain, M. S., Ahmmed, A., Begum, S., & Sadakuzzaman, M. (2024). Sensory, nutrition and physiochemical properties of beef in relation to their genotype. Meat Research, 4(3), Article 93. https://doi.org/10.55002/mr.4.3.93

Honikel, K. O. (1998). Reference methods for the assessment of physical characteristics of meat. Meat Science, 49, 447-457. https://doi.org/10.1016/S0309-1740(98)00034-5

Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. Statistics and Computing, 31, 82. https://doi.org/10.1007/s11222-021-10057-z

Hossain MA, Rahman MM, Rahman MW, Hossain MM, Hashem MA. 2023c. Effect of supplementary feeding on the production traits, carcass and meat quality of Jamuna basin lambs. Journal of Animal Science and Technology. 65 (1): 209-224.

Hossain MA, Rahman MM, Rahman MW, Hossain MM, Hashem MA. 2021b. Optimization of slaughter age of Jamuna basin lamb based on carcass traits and meat quality. SAARC Journal of Agriculture, 19(2): 257-270.

Huff-Lonergan, E., & Lonergan, S. M. (2005). Mechanisms of water-holding capacity of meat: The role of postmortem biochemical and structural changes. Meat Science, 71, 194-204. https://doi.org/10.1016/j.meatsci.2005.04.022

ISO 8586. (2023). Sensory analysis - Selection and training of sensory assessors. International Organization for Standardization, Geneva, Switzerland.

ISO 8589. (2007). Sensory analysis - General guidance for the design of test rooms. International Organization for Standardization, Geneva, Switzerland.

Kabir, M. H., Hashem, M. A., Ahmmed, A., & Ali, M. S. (2025). Effect of different concentration of moringa leaf extract (MLE) on proximate components, cooking loss, pH and color attributes of broiler meat sausage batter. Meat Research, 5(2), Article 116. https://doi.org/10.55002/mr.5.2.116

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lepetit, J. (2008). Collagen contribution to meat toughness: Theoretical aspects. Meat Science, 80, 960-967. https://doi.org/10.1016/j.meatsci.2008.06.016

Li, X., Ha, M., Warner, R. D., & Dunshea, F. R. (2022). Meta-analysis of the relationship between collagen characteristics and meat tenderness. Meat Science, 185, 108717. https://doi.org/10.1016/j.meatsci.2021.108717

Loomas, K. R., Woerner, D. R., Bohrer, B. M., Brown, T. R., Bruce, H. L., Duarte, M. S., Roy, B. C., Wang, Y., Pedgerachny, K., & Legako, J. F. (2025). Evaluation of rapid evaporative ionization mass spectrometry (REIMS) for the prediction of slice shear force and quality grades in beef longissimus lumborum steaks. Meat Science, 222, 109752. https://doi.org/10.1016/j.meatsci.2025.109752

Loomas, K. R., Woerner, D. R., Bohrer, B. M., Brown, T. R., Bruce, H. L., Duarte, M. S., Roy, B. C., Wang, Y., Pedgerachny, K., & Legako, J. F. (2026). Evaluation of rapid evaporative ionization mass spectrometry (REIMS) relationships with biochemical markers of tenderness in beef Longissimus lumborum steaks. Meat Science, 236, 110063. https://doi.org/10.1016/j.meatsci.2026.110063

Pearce, K. L., Rosenvold, K., Andersen, H. J., & Hopkins, D. L. (2011). Water distribution and mobility in meat during the conversion of muscle to meat and ageing and the impacts on fresh meat quality attributes - A review. Meat Science, 89, 111-124. https://doi.org/10.1016/j.meatsci.2011.04.007

Purslow, P. P. (2005). Intramuscular connective tissue and its role in meat quality. Meat Science, 70, 435-447. https://doi.org/10.1016/j.meatsci.2004.06.028

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. Nature Communications, 15, 1829. https://doi.org/10.1038/s41467-024-46150-w

Schroepfer, M., & Meyer, M. (2017). DSC investigation of bovine hide collagen at varying degrees of crosslinking and humidities. International Journal of Biological Macromolecules, 103, 120-128. https://doi.org/10.1016/j.ijbiomac.2017.04.124

Tornberg, E. (2005). Effects of heat on meat proteins - Implications on structure and quality of meat products. Meat Science, 70, 493-508. https://doi.org/10.1016/j.meatsci.2004.11.021

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. PLoS ONE, 14, e0224365. https://doi.org/10.1371/journal.pone.0224365

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

Wilson, M. D., Hunt, I. G., Sawyer, S. M., Kocharunchitt, C., Li, Z., & Stanley, R. A. (2025). Electronic tongue measurements as a predictor for sensory properties of vacuum-packed minced beef - A preliminary study. Meat Science, 220, 109705. https://doi.org/10.1016/j.meatsci.2024.109705

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

Zając, M., Pająk, P., & Skowyra, G. (2021). Characterization of edible collagen casings in comparison with the ovine casing and their effect on sausage quality. Journal of the Science of Food and Agriculture, 101, 6001-6009. https://doi.org/10.1002/jsfa.11254

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x