Research Square

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

A Novel Machine Learning-Based Approach for Adulteration Detection and Quantification in Surimi Using Physicochemical and Elemental Features

Linjun Zhou
Shanghai Ocean University
Lingyun Ma
Shanghai Ocean University
Xianhai Song
Shanghai Ocean University
Xinhui Yang
Shanghai Ocean University
Bian Li
Shanghai Ocean University
Hongmin Zhang
Shanghai Ocean University
Yong Zhao
Shanghai Ocean University
Qingchao Xie
qcxie@shou.edu.cn

Shanghai Ocean University

Research Article

Keywords: Surimi, fish category classification, adulteration detection, machine learning, SHAP

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11398037/v1

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

Additional Declarations: No competing interests reported.

A Novel Machine Learning-Based Approach for Adulteration Detection and Quantification in Surimi Using Physicochemical and Elemental Features

Linjun Zhou, Lingyun Ma[#], Xianhai Song, Xinhui Yang, Bian Li, Hongmin Zhang, Yong Zhao, Qingchao Xie [*]

# These authors have contributed equally to this work and share frst authorship.

∗ Corresponding author.

E-mail addresses: qcxie@shou.edu.cn (Q.c. Xie)

Abstract

Surimi adulteration poses a significant challenge to food safety. This study developed a machine learning framework that uses 13 physicochemical and elemental features for fish category classification and adulteration-ratio prediction. Four models (RF, XGBoost, LightGBM, and MLP), with logistic regression and ridge regression baselines, were evaluated using 30 source composition records, simulated mixtures and 30 measured validation samples. Five outer record folds and three inner folds separated model selection from evaluation. Principal component analysis (PCA) and SHAP were applied for feature analysis and model interpretation. RF and LightGBM each correctly classified all 30 held-out source records. XGBoost showed the lowest mean internal regression RMSE (0.1061; $R^2 = 0.8799$). In the 30 measured validation samples at 0%, 30% and 50%, RF gave the numerically lowest validation error (MAE = 0.0542, RMSE = 0.0791, $R^2 = 0.8517$), while the internally selected XGBoost gave MAE = 0.0619, RMSE = 0.0792 and $R^2 = 0.8515$. This measured set was not used for model fitting, preprocessing, model selection or noise selection, but its sampling breadth is limited. SHAP analysis identified fat, moisture and zinc as prominent classification features and fat, iron and sodium as prominent regression features. These results support exploratory composition-based modelling; broader independent measured validation is still required for practical regulatory applications.

Keywords: Surimi; fish category classification; adulteration detection; machine learning; SHAP

1 INTRODUCTION

Surimi products are widely used because of their nutritional value and processing adaptability, but processing removes many visible characteristics of the original fish, making species substitution difficult to recognise from appearance alone. A molecular survey of 56 surimi-based fish cakes sold through Chinese e-commerce found frequent disagreement between product labels and DNA-based species identification (X. Zhang et al., 2024). A separate Nanopore metabarcoding study of 20 commercial surimi products also detected complex multispecies composition and product misrepresentation (Detcharoen et al., 2024). These findings demonstrate the practical need for reliable surimi-authentication tools because substitution can affect product quality, label compliance and consumer choice.

Current surimi-authentication methods provide different forms of evidence. DNA barcoding is an established approach for confirming seafood species identity (Filonzi et al., 2023), whereas next-generation sequencing can resolve multiple taxa in mixed processed seafood products (Piredda et al., 2022). The composition-based strategy evaluated here is complementary: it uses routinely interpretable physicochemical and elemental concentrations to classify source categories and estimate an adulterant fraction. It is therefore intended as an additional analytical option rather than a replacement for molecular confirmation.

Fish species exhibit interspecific differences in nutritional composition and mineral-element accumulation. Physicochemical nutritional indicators—including protein, fat, moisture, and ash contents—reflect the biochemical composition of fish materials. Mineral elements can vary with metabolic characteristics, feeding behaviour, and habitat environment. Reviews of fish and seafood authentication show that multi-element and stable-isotope profiles can support origin, production-method and species discrimination (Han et al., 2022; Varrà et al., 2021), and the broader food-authentication literature likewise identifies multi-element chemometrics as a useful analytical fingerprint (Zaldarriaga Heredia et al., 2023). These data retain interpretable concentration units, although database definitions and processing also influence the recorded values.

Machine learning provides several suitable approaches for learning nonlinear patterns in tabular composition data. Random forest (RF) averages randomised decision trees (Breiman, 2001), whereas extreme gradient boosting (XGBoost) (Chen & Guestrin, 2016) and light gradient boosting machine (LightGBM) (Ke et al., 2017) build additive boosted-tree models with different optimisation and tree-growth strategies. A multilayer perceptron (MLP) offers a neural-network comparator. Combining these models with physicochemical and elemental features enables both qualitative identification and quantitative prediction to be tested within one analytical framework.

The overall workflow of this study comprises two complementary analytical tasks that jointly serve the goal of surimi adulteration detection. The first task is fish category classification, which aims to distinguish pure cod, tilapia and catfish samples based on their physicochemical

and elemental composition. The second task is adulteration-ratio prediction, which aims to quantify the proportion of an adulterant species (tilapia or catfish) in a mixed surimi sample. Machine learning framework for qualitative classification and quantitative prediction that integrates multidimensional chemical-composition features. Recent reviews show that chemometrics and machine learning are increasingly integrated with rapid food-authenticity measurements (Goyal et al., 2024; Z. Zhang et al., 2024), while fish-species studies using multimode spectroscopy and fused LIBS-Raman signals demonstrate the value of comparing multiple nonlinear classifiers (Ren et al., 2023; Sueker et al., 2023). Four models (RF, XGBoost, LightGBM, and MLP) and linear baselines were established using nested cross-validation. Classification distinguishes the cod, tilapia and catfish source categories, whereas a separately trained common regressor predicts adulterant mass fraction for the two binary mixing series. SHAP interpretability analysis was employed to examine the contribution of key chemical features. Tree ensembles and an MLP were selected to compare bagging, boosting and neural-network learning in a small tabular problem. This study provides a comparative application of existing algorithms, without an algorithmic modification, and evaluates their behaviour under stated data and simulation assumptions. By addressing both qualitative and quantitative aspects, this framework offers a more complete analytical solution for surimi authenticity verification than either task alone.

2 MATERIALS AND METHODS

2.1 Data sources and sample composition

The physicochemical and elemental-composition data used in this study were collated from public food-composition databases, including the United States Department of Agriculture (USDA), the Food and Agriculture Organization of the United Nations (FAO), and the food-nutrition composition query platform published by the National Institute for Nutrition and Health (Chinese Center for Disease Control and Prevention) and the Chinese Nutrition Society. The specific database versions, parent food items or food identifiers, and source identifiers for the 30 records are provided in Supplementary Table S1. Cod, tilapia, and catfish were selected as research subjects. Cod served as the primary species for pure-surimi sample construction, whereas tilapia and catfish were used as adulterant species for composition-ratio analysis (Fayaz et al., 2021; González et al., 2020).

The dataset represents 30 food-composition records rather than 30 verified independent fish. The different record types and shared underlying sources limit the interpretation of biological variation and record-level validation. Database labels and purchasing regions do not establish catch origin. The original predictor values were retained for the reported analysis.

The original data comprised physicochemical nutritional indicators and mineral-element indicators. The physicochemical indicators included fat, protein, moisture, and ash contents.

The mineral-element indicators included calcium (Ca), iron (Fe), magnesium (Mg), phosphorus (P), potassium (K), sodium (Na), zinc (Zn), copper (Cu), and selenium (Se). These indicators reflect differences among source categories in basic nutritional composition and element-accumulation characteristics. They thereby provide multidimensional chemical-composition information for surimi authenticity research.

During preprocessing, the data formats and measurement units of all indicators were checked for consistency before subsequent analysis. The original dataset comprised 30 composition records across three fish categories: 8 cod, 12 tilapia, and 10 catfish. No missing values were present in any of the 13 retained physicochemical or mineral-element variables, so no rows were removed or imputed.

Because publicly available fish-composition data are limited, this study used virtual sample generation for the regression task. Such strategies are designed to increase coverage of the input-output space when measured samples are scarce (Wen et al., 2025). The adulteration-ratio gradient was 0%, 10%, 20%, ..., 100% (11 levels); 120 pure-cod rows and 60 mixed rows per nonzero level were generated for each of two adulterants; and 5% relative Gaussian perturbation was applied. Each source partition therefore yielded 1,320 virtual rows: 120 pure cod, 600 cod-tilapia and 600 cod-catfish rows.

2.2 Physicochemical and elemental features

Physicochemical nutritional indicators and mineral-element indicators of fish samples were selected as model inputs to construct the surimi-composition analysis dataset. The physicochemical indicators included protein, fat, moisture, and ash contents. The mineral-element indicators included Ca, Fe, Mg, P, K, Na, Zn, Cu, and Se.

The 13 predictors comprise four proximate indicators in g/100 g of edible portion, eight elements (Ca, Fe, Mg, P, K, Na, Zn and Cu) in mg/100 g, and Se in μg/100 g. Carbohydrate was excluded before modelling because it may be calculated by difference and its definition is not consistently comparable across food-composition databases, a known data-harmonisation concern (Charrondiere et al., 2016). It is excluded from all model inputs, composition figures, PCA and SHAP results; the complete original workbook is retained for traceability.

Physicochemical composition reflects fundamental differences in fish materials, while elemental profiles provide an additional and complementary description. Elemental analysis combined with chemometrics has differentiated salmonid production methods and origins (Han et al., 2021), demonstrating that multielement patterns can carry authentication information. Processing can also influence the final composition of fish products (Sampels, 2015). The present study therefore interprets the measured concentrations as an integrated analytical fingerprint rather than assigning each observed difference to a single biological cause.

By integrating physicochemical nutritional indicators with mineral-element features, this study constructed a surimi chemical-feature system with multidimensional composition information. This integration provides feature information for surimi composition analysis and authenticity identification.

2.3 Data augmentation and mixed-sample simulation

Given the limited number of source composition records, a virtual-sample strategy was used to represent the full adulteration-ratio range. Virtual sample generation is an established small-sample modelling approach for expanding coverage between observed data points (Wen et al., 2025). Here, the procedure was defined by the physical mass fractions of cod and the adulterant and was applied only to regression.

During the mixed-sample construction process, cod was designated as the primary fish species and separately mixed with tilapia and catfish at varying composition ratios. The feature values of the mixed samples were constructed using linear weighting, calculated as follows:

$$z_j = (1 - r)c_j + ra_j \tag{1}$$

where z_j is feature j of the mixture before perturbation; c_j and a_j are the corresponding features of cod and the adulterant species, respectively; and r is the adulterant mass fraction (0–1). The index j identifies one of the 13 retained composition features.

To examine model behaviour under composition fluctuations, a Gaussian perturbation was introduced during the mixed-sample generation process. The Gaussian noise level was set at 5% (relative standard deviation). This is a modelling assumption, not a parameter estimated from individual biological or analytical variation. Independent feature noise does not preserve empirical covariance, compositional closure or process-specific losses. The perturbed sample features are expressed as follows:

$$x_j = max\{0,\ z_j + \sigma|z_j|\varepsilon_j\}, \quad \varepsilon_j \sim N(0,1) \tag{2}$$

where x_j is the generated feature after perturbation and ε_j is an independent standard normal random variable. The relative noise level is $\sigma = 0.05$ in the primary analysis. Negative generated concentrations are set to zero; a zero-valued source component remains zero under this perturbation.

To prevent direct source-record leakage between the training and test sets, this study first partitioned the source records and subsequently constructed mixed-sample datasets independently for the training and test sets. Specifically, stratified 5-fold cross-validation (StratifiedKFold, random_state = 20260908) was employed, whereby the original records in each fold were stratified by fish category into 24 training and 6 test records. Parameter selection used three stratified inner folds of the current training records. Virtual samples were

independently generated after each inner or outer split, ensuring that the corresponding held-out records did not participate in simulated training for that evaluation.

Within each partition, eligible parents were selected by randomly ordered balanced cycling. The total was $120 + 2 \times 10 \times 60 = 1{,}320$ rows, including 120 rows at every target fraction. The zero-fraction group was generated once, rather than duplicated across the two series. Five outer test folds produced 6,600 correlated simulated predictions per model. Full-source fitting used 1,320 generated training rows. Classification used only the 30 unaugmented records.

Record partitioning does not guarantee independence of the underlying studies. FAO records 091002 and 091003 cite 091001 and SR27-15261; their assigned source rows 18–20 occur on opposite sides of retained outer folds 2 and 4. The reported evaluation is therefore a record-level comparison, not validation on independent sampling studies. The generated training and test data also share the same mixing and noise assumptions (Kulkarni & Goswami, 2015; Lee et al., 2010).

2.4 Machine learning model construction

2.4.1 Classification model construction

The architectures of the five evaluated models are schematically shown in Fig. 1. To achieve qualitative identification of fish composition types, RF, XGBoost, LightGBM and MLP classifiers were constructed, with logistic regression as a linear baseline. These models represent bagging, gradient boosting and neural-network learning (Breiman, 2001; Chen & Guestrin, 2016; Ke et al., 2017). The labels were cod, tilapia and catfish for the original pure-material records; no mixed-sample class or adulteration-ratio label was used. Tree models received original-unit features. MLP and logistic regression used StandardScaler fitted only within each training pipeline. Stratified cross-validation was used to maintain category balance.

2.4.2 Regression model construction

To achieve quantitative prediction of surimi composition ratios, RF, XGBoost, LightGBM and MLP regressors were constructed, with ridge regression as a linear baseline. Model inputs comprised the 13 physicochemical and elemental features, and the output was the adulterant-species mass fraction (0-1). The cod-tilapia and cod-catfish series were pooled to train one common regressor per algorithm. The target was 0 for pure cod, the known mass fraction for partial substitution and 1 for complete replacement. MLP and ridge regression used training-only standardisation; tree models received the original units.

The classifier and regressor were trained and evaluated separately. The regressor received neither true species nor predicted category, mixture-series label or known target fraction as an input. A cod classification did not force a regression prediction to zero, and classification did not select between separate ratio regressors. Shared source compositions do not imply shared task labels.

2.4.3 Hyperparameter optimisation

A nested hyperparameter-selection strategy was adopted to keep model selection separate from outer-fold evaluation, following established guidance on avoiding selection bias in performance assessment (Cawley & Talbot, 2010). In each outer training partition, three prespecified configurations per model and task were compared using 3-fold inner cross-validation. Accuracy was the optimisation objective for classification, with macro-F1 breaking ties; RMSE was the objective for regression, with MAE breaking ties. Remaining ties followed candidate-list order. Five outer folds evaluated the selected configurations. A final three-fold source-only search selected parameters for full-source fitting after outer evaluation.

Hyperparameter configurations covered tree counts and structural constraints for RF, XGBoost and LightGBM, hidden widths of 10, 20 or 32 ReLU units for MLP, and regularisation for the linear baselines. The classification MLP used L-BFGS; the regression MLP used Adam with a 2,000-iteration maximum and early stopping on 15% of its current training descendants. That stopping subset may share source parents with other training descendants and is not an independent source validation. All final 13-predictor fits converged. Implementations used Python 3.13.3, scikit-learn 1.6.1, XGBoost 3.0.1 and LightGBM 4.6.0.

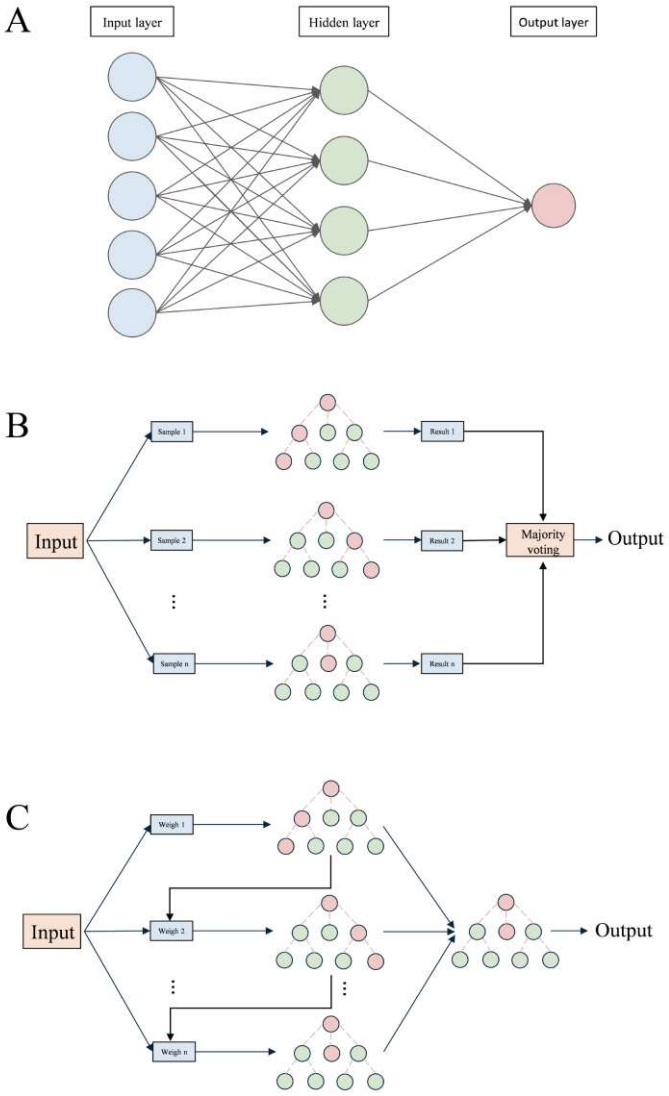


Figure 1. Schematic diagrams of three types of machine learning model architectures: (A) neural network (NN) structure; (B) bagging ensemble method; (C) boosting ensemble method.

2.5 Model evaluation metrics

2.5.1 Classification model evaluation metrics

Classification performance was evaluated using accuracy, precision, recall and F1-score. Accuracy measures the proportion of correct predictions. Precision describes the correctness of predictions assigned to a category, recall describes the proportion of a category that is recovered, and F1-score is the harmonic mean of precision and recall. Precision, recall and F1 were macro-averaged so that the three fish categories received equal weight.

The formulas for accuracy, precision, recall, and F1-score are as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively. Equation (3) gives binary accuracy. For the three-category results, overall accuracy is the number of correctly classified records divided by the number evaluated. Precision, recall and F1 were calculated for each class against the remaining classes, then macro-averaged with equal class weights. Class metrics with a zero denominator were set to zero before averaging. Each source record received one outer-fold prediction per model.

2.5.2 Regression model evaluation metrics

To evaluate the prediction performance of regression models for surimi composition ratios, the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were adopted as model evaluation metrics.

Among these, R^2 measures the goodness of fit between model-predicted values and actual values; RMSE summarises the magnitude of model prediction errors; and MAE measures the average absolute deviation between predicted and actual values. Higher R^2 values and lower RMSE and MAE values indicate better model prediction performance (Qu et al., 2024).

The formulas for the evaluation metrics are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (9)$$

where y_i is the actual adulterant mass fraction, $\hat{y}_i$ is its predicted value, $\overline{y}$ is the mean of the actual values, and n is the number of evaluated rows.

Internal results are means and sample standard deviations across the five outer folds. Fold dispersion is descriptive and is not a population confidence interval because training sets overlap and simulated rows share parents. Predictions are not clipped to [0, 1]. An MAE of 0.01 equals one percentage point of ratio error. The measured validation set was evaluated once as

a complete 30-row set, and only pooled metrics for all 30 rows are reported. R^2 is undefined within a constant-target group.

2.6 SHAP interpretability analysis

Shapley Additive exPlanations (SHAP) was used to interpret the fitted models. SHAP is an additive feature-attribution framework that assigns a contribution value to each input for a particular prediction (Lundberg & Lee, 2017). This makes it possible to examine both the relative importance and the direction of contribution of physicochemical and elemental features.

In this study, SHAP analysis was primarily used to evaluate the influence of physicochemical nutritional indicators and mineral-element features on surimi composition classification and adulteration ratio prediction. By calculating the SHAP values of each feature, the direction and relative importance of contributions from different composition features to model predictions could be examined. Classification and regression were explained separately using the full-source RF classifier and XGBoost common regressor, respectively, selected from the internal evaluation. RF represents the tie with LightGBM rather than a uniquely superior classifier.

Tree SHAP used original-unit concentrations and tree-path-dependent reference statistics. The classifier was explained on the 30 source records; the regressor was explained on 220 sampled full-source training descendants. These are fitted-model explanations, not additional held-out evaluations. Classification SHAP values are probability contributions and regression SHAP values are mass-fraction contributions. Dependence axes retain the physical feature units; SHAP values are not inverse-standardised into nutrient concentrations. Maximum additivity residuals were 3.33×10^{-16} and 5.36×10^{-7}, respectively. Colouring by another feature indicates an exploratory association, not a formal interaction test.

2.7 Measured-sample documentation and quantitative evaluation

To evaluate quantitative prediction on measured material, 30 laboratory composition profiles were used: 10 pure-cod samples, 10 cod-tilapia mixtures at 30% adulteration and 10 cod-catfish mixtures at 50% adulteration. The profiles were taken from the recorded 30-row validation workbook. They were not augmented and were not used for model fitting, preprocessing fitting, hyperparameter selection or noise selection. Thus, this set provides a direct measured-data evaluation of models developed from the database records and virtual mixtures.

Raw fish materials were purchased in April 2026 through commercial retail channels. Cod was purchased from JD.com under Russian-cod and Alaska-cod commercial labels; tilapia was purchased as a three-part-removed product from Guangdong; and catfish was purchased as wild local catfish from Weishan. Approximately 10–15 individuals of each species were obtained. Species identity was based on commercial labelling and was not independently confirmed by molecular methods. Edible muscle portions were separated, minced and homogenised

separately for each species. For mixture preparation, the homogenised cod and adulterant tissues were blended at target mass ratios of 70:30 (cod:tilapia) and 50:50 (cod:catfish) by weight, then re-homogenised to ensure uniformity. Each group comprised 10 independent mixture preparations. Pure-cod samples were prepared from homogenised cod tissue without adulterant addition.

The measured workbook contains recorded feature values and nominal mixture ratios. The mixture proportions are reported as 0, 0.30 and 0.50 in the evaluation table; no generated blending or additional Gaussian perturbation was applied to these rows during evaluation.

Physicochemical indicators were determined according to national food-safety standards: moisture by GB 5009.3-2016 (direct drying, 105 °C), ash by GB 5009.4-2016 (dry ashing, 550 °C), protein by GB 5009.5-2025 using the Kjeldahl method (N × 6.25; UDK 159, VELP Scientifica, Italy), and fat by GB 5009.6-2025 using the Soxhlet extraction method (SZF-06C, Shanghai Jinghai, China). Mineral elements (Ca, Fe, Mg, P, K, Na, Zn, Cu, Se) were determined according to GB 5009.268-2025, Part I Method I, using ICP-MS (iCAP RQ; Thermo Fisher Scientific, USA). Sample digestion followed the procedure specified in GB 5009.268-2025: approximately 0.2–0.5 g of homogenised sample was digested with nitric acid and hydrogen peroxide in a microwave digestion system, then diluted to a fixed volume with ultrapure water. Germanium (Ge), rhodium (Rh) and bismuth (Bi) were used as online internal standards. Measurements occurred after March 2026.

Each sample was analysed in triplicate, and reported values are means. Replicate RSDs were below 3% for moisture and ash, below 5% for protein and fat. For ICP-MS, the calibration correlation coefficient (R^2) was ≥ 0.999 for all elements; LODs ranged from 0.001 to 0.010 mg/kg and LOQs from 0.005 to 0.030 mg/kg; spike recoveries ranged from 83% to 115% with RSDs below 5%. Method blanks, parallel samples and certified reference materials were included in each analytical batch.

The measured samples were kept separate from the source-record simulation and were evaluated once per row. The 30 recorded profiles were input directly into the trained regression models without retraining, clipping predictions or overwriting cod predictions. Each row was evaluated once, and no augmented copy of a measured sample was used. These results constitute the measured validation reported below.

2.8 Noise sensitivity analysis

RF was specified as the representative model for quantitative noise sensitivity at $\sigma = 0$, 0.02, 0.05 and 0.10. Each outer fold retained the parameters selected under the main 5% condition. Source partitions, parent assignments, ratio counts and underlying standard-normal draws were unchanged. Training and simulated test data used the same σ within each condition, so this is a

matched-simulator comparison. The 5% result was reused from the main analysis. Fish category classification was not augmented and therefore did not use this mixing-noise experiment.

3 RESULTS AND DISCUSSION

3.1 Compositional feature analysis of original data

Descriptive analysis of the 30 source records showed clear differences among cod (COD), tilapia (TIL) and catfish (CAT) in both proximate composition and elemental profiles (Fig. 2). This combination is analytically meaningful: lipid- and protein-related information has enabled discrimination among marine surimi types (Liu et al., 2015), while multielement profiles have supported chemometric authentication of salmonids (Han et al., 2021). The present data therefore bring two established sources of discriminatory information into one concentration-based feature set.

Among the proximate indicators, CAT records had the highest fat content and COD records the lowest, while TIL records showed relatively high protein content. Earlier infrared work likewise identified lipid- and protein-related information as useful for distinguishing three marine surimi materials (Liu et al., 2015). Although spectral variables and directly measured concentrations are not numerically equivalent, both findings indicate that proximate composition carries source-related information. The present records were not designed to test effects of habitat, diet or physiology, so the observed differences should be interpreted as analytical contrasts rather than evidence of a specific biological cause.

COD records had the highest moisture content overall, whereas CAT records had relatively lower moisture content (Fig. 2); the higher fat level of CAT coincided with a lower moisture proportion. Because fat and moisture fractions are compositionally related, their joint contribution is best viewed as a multivariate contrast rather than as two independent species markers. COD also showed slightly higher ash content than CAT, although this difference was less pronounced than the fat and moisture patterns.

The elemental results added information beyond proximate composition. COD records showed higher Zn and Se, TIL records relatively high Mg, and CAT records variation in Fe and Cu; the categories also differed in Na, K and P (Fig. 2). Elemental profiles combined with multivariate models have discriminated salmonid production methods and origins (Han et al., 2021), and reviews identify inorganic composition as a useful but context-dependent source of fish-authentication information (Varrà et al., 2021; Han et al., 2022). The present result applies this analytical principle to cod, tilapia and catfish records, but does not establish any individual element as a universal species marker because elemental concentrations can also vary with origin, environment and processing.

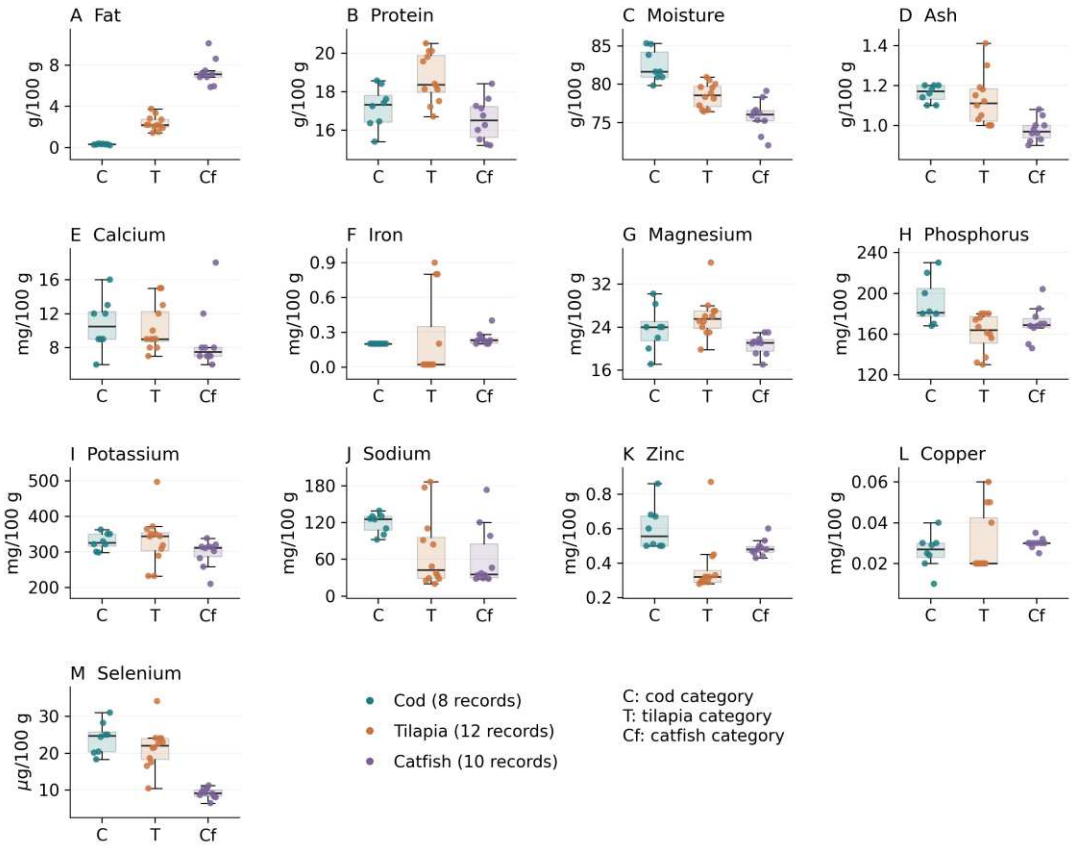


Figure 2. Distributions of all thirteen retained physicochemical and elemental predictors in the thirty source records (8 cod, 12 tilapia and 10 catfish). Units are shown on the individual panels. Points are recorded values; boxes show the median and interquartile range, with whiskers extending to the most extreme observations within 1.5 interquartile ranges of the box. The distributions are descriptive and contain no inferential significance annotations.

3.2 Principal component analysis

To visualise distribution patterns of different species in the multidimensional feature space and examine compositional-gradient variations of mixed samples, principal component analysis (PCA) was employed. All features were standardised before extracting the first two principal components.

StandardScaler and PCA were fitted to the 30 source records using the same 13 predictors. The 1,320 saved full-source training mixtures were projected using this fixed scaler and component basis. This descriptive whole-library analysis was separate from model selection and does not replace training-only preprocessing in the predictive pipelines.

PC1 and PC2 explained 28.95% and 23.25% of the variance, respectively, for a combined 52.19% (Fig. 3A). Cod, tilapia and catfish occupied partly distinct regions, with overlap mainly between some COD and TIL records. The grouping suggests that the 13 features contain a

multivariate source-category signal, while the overlap and the 47.81% of variance outside the first two components show that the two-dimensional projection is not a complete separation test. This qualitative pattern is consistent with the general feasibility of multivariate surimi discrimination (Liu et al., 2015). Fig. 3(B) shows the projected distribution of the virtual mixtures. As the adulteration ratio increased from 0% (pure cod) to 100% (pure adulterant), the cod-tilapia and cod-catfish series formed continuous composition trajectories from the pure-cod region toward their respective adulterant regions. This orderly gradient confirms that the mass-weighted generation procedure represents the intended ratio progression in multivariate feature space, providing a coherent basis for subsequent regression modelling.

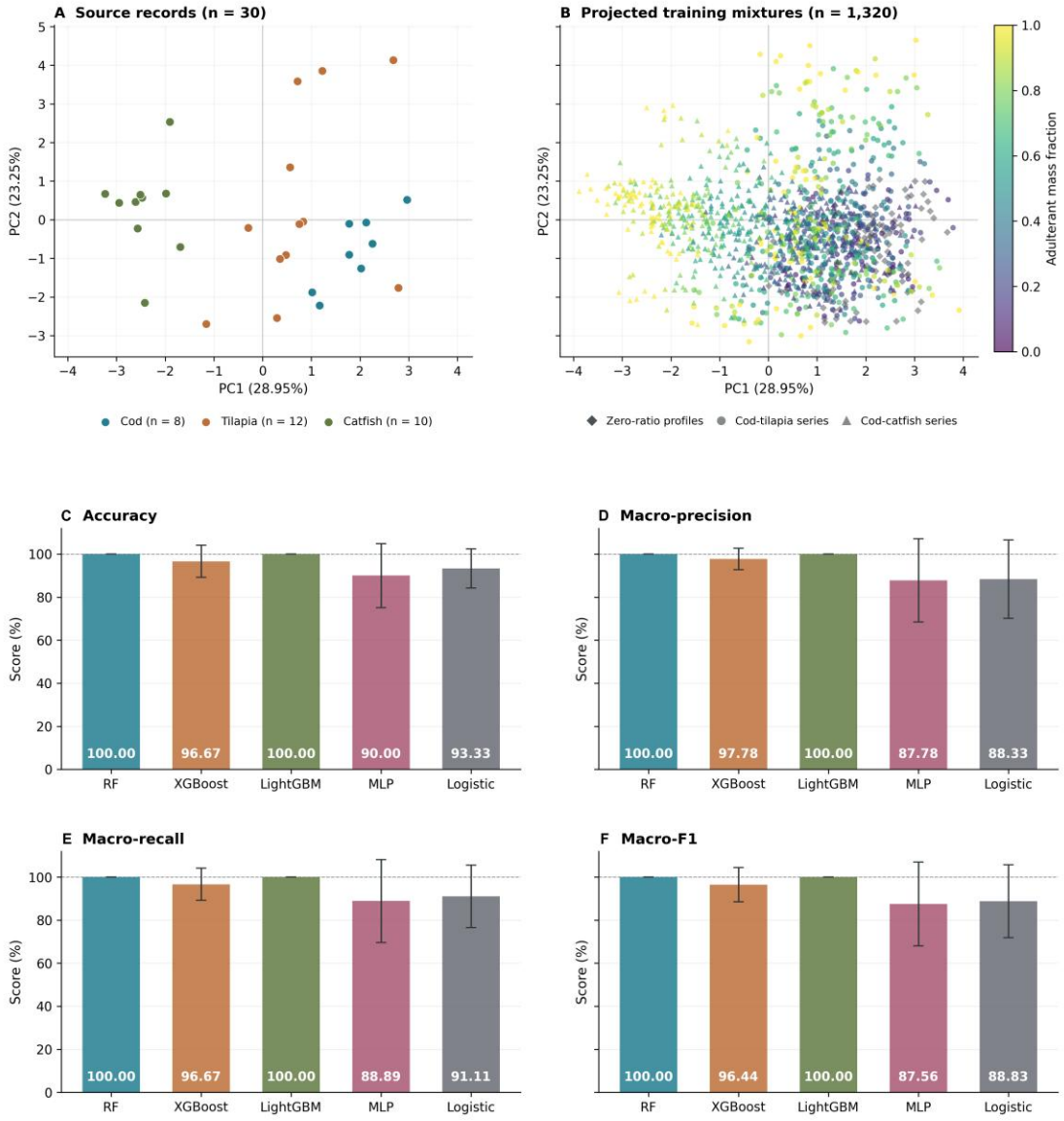


Figure 3. Descriptive PCA and classifier-performance results. Panels A and B show the thirty source records (8 cod, 12 tilapia and 10 catfish) and the 1,320 saved training mixtures projected using the scaler and principal components fitted only to the source library, respectively. PC1 and PC2 explain 28.95% and 23.25% of source variance. The projection is exploratory and was not used to select or score prediction models. Panels C-F show accuracy, macro precision, macro recall and macro F1, respectively, for all five thirteen-predictor classifiers. Bars and error bars show five-fold means and sample standard deviations and correspond to Table 1. Fold dispersion is descriptive rather than a population confidence interval.

3.3 Classification analysis using machine learning models

To achieve qualitative identification of the source fish categories, four classification models—RF, XGBoost, LightGBM, and MLP—and logistic regression were constructed. All models used 13 physicochemical and mineral-element features as inputs. Models were trained and their

hyperparameters were selected using stratified 5-fold outer cross-validation with 3-fold inner selection; standardisation was applied only where specified in Section 2.4.

RF and LightGBM each classified all 30 held-out source records correctly; XGBoost correctly classified 29, logistic regression 28 and MLP 27 (Table 1). The agreement across nonlinear and linear models indicates that the combined physicochemical-elemental profile was strongly discriminative within this source library. Machine-learning-based fish and surimi classification has also been demonstrated with infrared fingerprints (Liu et al., 2015), multimode spectroscopy (Sueker et al., 2023) and fused LIBS-Raman signals (Ren et al., 2023). These platforms use different analytical signals, sample sets and validation designs, so their reported accuracies are not direct benchmarks for the present 30-record nested cross-validation. In particular, the perfect RF and LightGBM results should be interpreted as internal held-out performance within the defined three-category library, not as evidence of error-free classification of independent commercial mixtures.

Table 1. Nested five-fold classification of the 30 unaugmented source records.

Model	Accuracy (%)	Macro precision	Macro recall	Macro-F1
RF	100.00 ± 0.00	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
XGBoost	96.67 ± 7.45	0.9778 ± 0.0497	0.9667 ± 0.0745	0.9644 ± 0.0795
LightGBM	100.00 ± 0.00	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
MLP	90.00 ± 14.91	0.8778 ± 0.1940	0.8889 ± 0.1925	0.8756 ± 0.1947
Logistic regression	93.33 ± 9.13	0.8833 ± 0.1826	0.9111 ± 0.1449	0.8883 ± 0.1690

Entries are mean ± sample SD across five outer folds. Logistic regression is the linear classification baseline. Pooled confusion counts use exactly one prediction per source record; macro fold averages need not equal macro metrics calculated after pooling all predictions.

Fig. 3C-F compares accuracy, macro precision, macro recall and macro F1 across the five classifiers. RF and LightGBM reached the maximum value for all four metrics, while XGBoost, logistic regression and MLP also maintained high overall performance. The error bars show the observed fold-to-fold variation and support the consistency of the leading models.

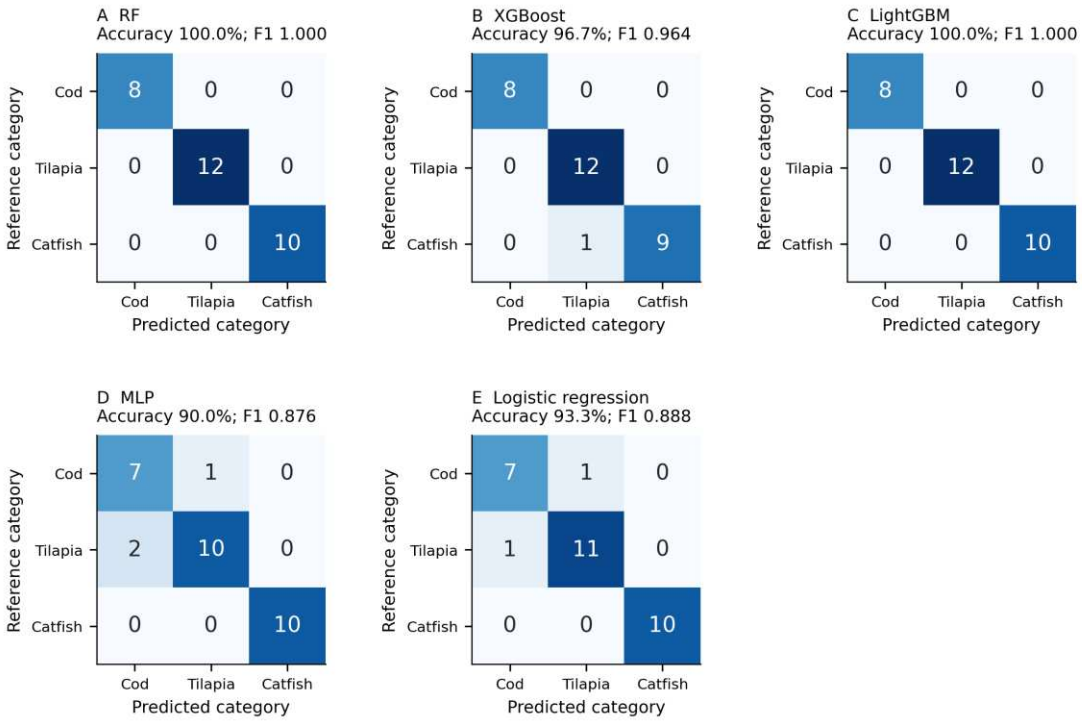


Figure 4. Complete confusion matrices for RF, XGBoost, LightGBM, MLP and logistic regression. Each panel contains one held-out prediction for each of thirty source records, with categories ordered as cod, tilapia and catfish. Accuracy and macro-F1 annotations are outer-fold means matching Table 1. These matrices concern source-record labels, not external classification of mixed products.

Fig. 4 presents the confusion matrices on the test folds. Panels correspond to RF, XGBoost, LightGBM, MLP and logistic regression, respectively. RF and LightGBM made no classification errors. XGBoost misclassified one catfish record as tilapia. MLP misclassified one cod record as tilapia and two tilapia records as cod; logistic regression confused one cod and one tilapia record. The matrices use exactly one held-out prediction per original record and concern the three source categories, not adulteration classes or external mixture identification.

3.4 Regression analysis for adulteration ratio prediction

To achieve quantitative prediction of adulterant-species ratios in surimi, four regression models—RF, XGBoost, LightGBM, and MLP—and ridge regression were constructed. Model inputs comprised 13 physicochemical and mineral-element features. The output was the corresponding adulterant-species ratio (0–1). Both mixing series were pooled for each common regressor. Hyperparameters for all models were selected in the inner folds, and the models were evaluated through 5-fold outer record-level cross-validation.

Fig. 5 presents the true-versus-predicted adulterant fractions for all five regressors. XGBoost achieved the lowest mean internal RMSE (0.1061; R-squared = 0.8799; MAE = 0.0804), while RF achieved the lowest MAE (0.0774). LightGBM and RF produced similarly strong R-squared

values of 0.8750 and 0.8751, respectively. The agreement among the three tree families indicates that ratio information was not dependent on one ensemble implementation. Quantitative surimi-authenticity studies have used infrared spectroscopy for starch (Hou et al., 2019), RGB-derived reconstructed spectra (Feng et al., 2023), volatile fingerprints for fish-mixture discrimination (Chen et al., 2024), and FT-NIR for adulterant classification and fraction prediction in Alaska pollock (Feng et al., 2024). (Feng et al. 2024) reported an SVR prediction R-squared of 0.946 and an RMSEP of 0.062, while (Feng et al. 2023) reported prediction correlations above 0.91 and RMSE values below 9% for most reconstructed-spectral models. These results place the present performance within the range demonstrating practical quantitative feasibility, but they are not direct rankings: the adulterants, analytical signals, ratio designs, sample sizes and validation partitions differ, and the present estimates derive from source-partitioned nested evaluation of composition-based virtual mixtures.

Table 2. Common-regressor performance on held-out synthetic mixtures with 13 predictors.

Model	MAE	RMSE	R^2
RF	0.0774 ± 0.0120	0.1107 ± 0.0174	0.8751 ± 0.0370
XGBoost	0.0804 ± 0.0191	0.1061 ± 0.0309	0.8799 ± 0.0719
LightGBM	0.0821 ± 0.0193	0.1086 ± 0.0297	0.8750 ± 0.0682
MLP	0.1207 ± 0.0233	0.1549 ± 0.0330	0.7514 ± 0.0969
Ridge regression	0.1920 ± 0.0434	0.2399 ± 0.0543	0.4007 ± 0.2530

Mean ± sample SD over five source folds, with 1,320 generated test rows per fold. All 11 ratio levels, including zero, contribute. Errors are on the 0–1 scale. The 6,600 generated rows are correlated descendants of 30 source records, not 6,600 independently measured samples.

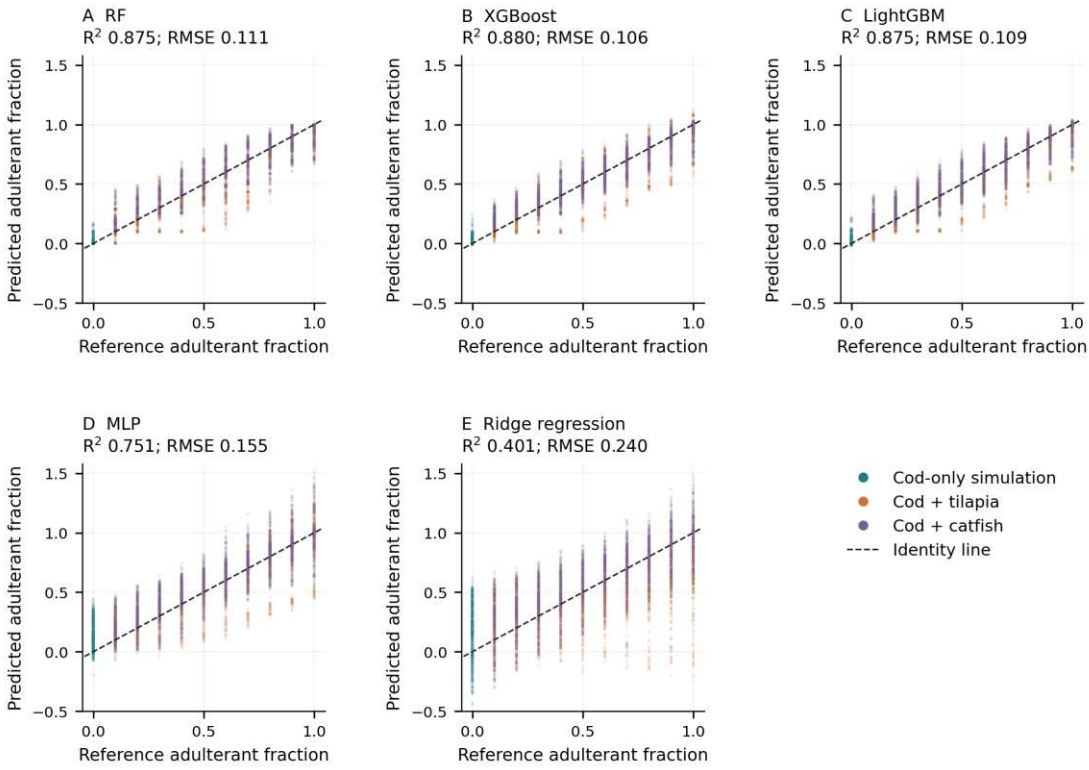


Figure 5. True versus predicted adulterant fractions for all five thirteen-predictor regressors in internal evaluation. Each panel uses 6,600 correlated generated test descendants across five record folds. Annotated RMSE and R^2 are fold means matching Table 2, not metrics recomputed after pooling folds. The dashed line denotes identity; predictions are untruncated.

3.5 SHAP analysis

To enhance model interpretability and reveal the contributions of key physicochemical and elemental features to prediction results, SHAP (Shapley Additive exPlanations) was employed for interpretability analysis of the selected models. RF was used as a representative of the tied classification models, whereas XGBoost was selected by the internal regression RMSE criterion. Classification and regression explanations were calculated separately.

The largest classification contributions were associated with fat, moisture and zinc, whereas the largest regression contributions were associated with fat, iron and sodium (Fig. 6A and 6D). The prominence of fat is compatible with the lipid-related differentiation of surimi reported previously (Liu et al., 2015), and the contribution of several elements is consistent with the discriminatory value of multielement fish profiles (Han et al., 2021). Fat and moisture primarily helped distinguish the overall composition of the source categories, whereas regression importance also reflected variables that changed systematically along the simulated mixture trajectories. Proximate and elemental variables therefore supplied complementary model information rather than one feature family dominating both tasks.

The SHAP beeswarm plots further show the direction and distribution of these contributions. Higher fat generally increased the predicted adulteration fraction (Fig. 6E), whereas iron showed a nonlinear pattern across mixtures (Fig. 6F). For the cod-classification output, lower fat and higher moisture generally supported the cod prediction (Fig. 6B-C). Classification asks which source category a profile resembles, whereas regression asks how features vary with mixture fraction; different rankings are therefore expected even when the same predictors are used. These SHAP values describe contributions within the fitted models and should not be interpreted as causal effects or as validation of fat, iron, sodium or zinc as universal biomarkers.

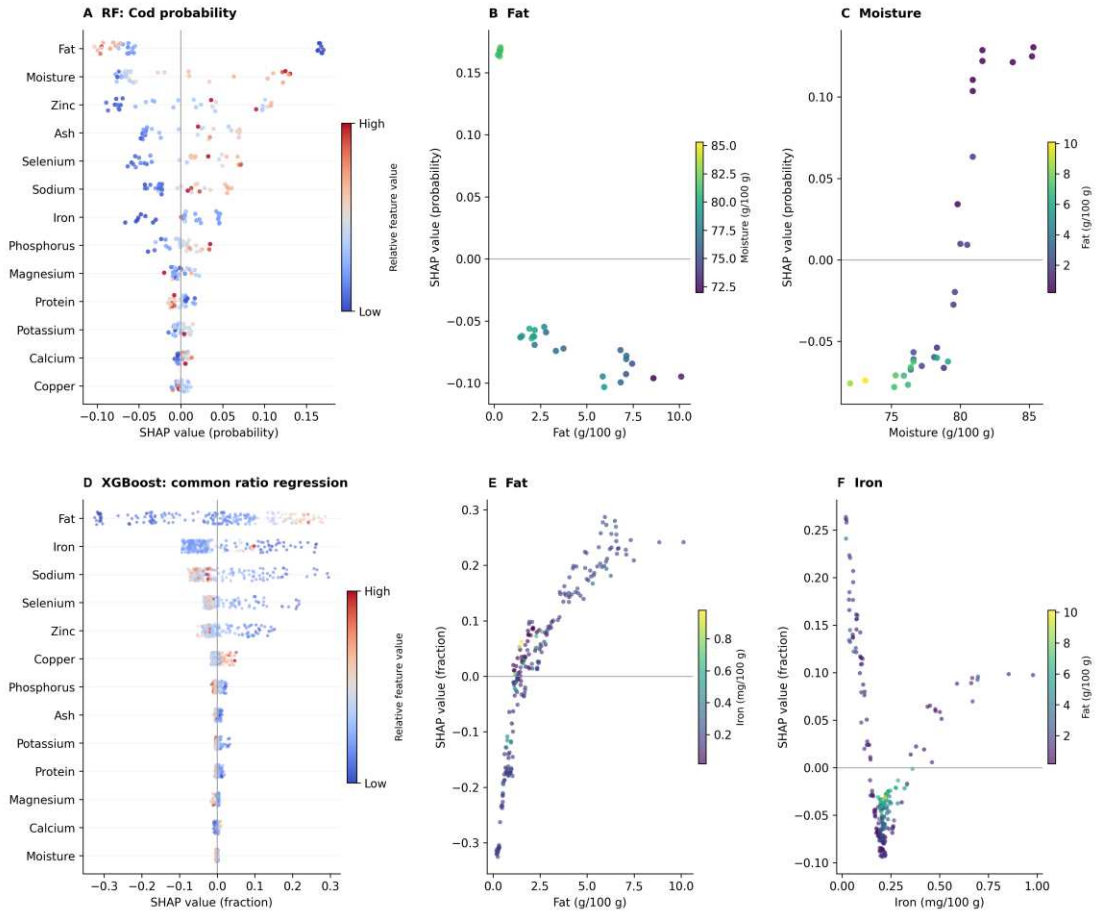


Figure 6. SHAP interpretation of the thirteen-predictor models. Panels A-C show the RF classification SHAP summary (A) and dependence plots for fat (B) and moisture (C) for the cod probability output; summary colours indicate relative feature values, dependence axes retain original concentration units, and SHAP values are probability contributions. Panels D-F show the XGBoost regression SHAP summary (D) and dependence plots for fat (E) and iron (F); SHAP values contribute to the predicted adulterant mass fraction, and dependence axes retain original units. These plots explain the fitted simulation models rather than establish causality or independently measured accuracy.

All dependence axes retain original concentration units; SHAP output units follow Section 2.6.

3.6 Measured quantitative validation results

To assess quantitative prediction on real measured samples, 30 measured profiles were evaluated: 10 pure-cod samples, 10 cod–tilapia mixtures at 30% adulteration, and 10 cod–catfish mixtures at 50% adulteration. They were input directly into the frozen common regression models without refitting or augmentation.

On the complete set of 30 measured samples (Fig. 7A-E; Table 3), RF gave MAE = 0.0542, RMSE = 0.0791 and R-squared = 0.8517. The internally selected XGBoost gave MAE = 0.0619, RMSE = 0.0792 and R-squared = 0.8515, while LightGBM gave MAE = 0.0626, RMSE = 0.0794 and R-squared = 0.8507. The RMSE values differed by only 0.0003 across the three tree models; without repeated external validation or uncertainty intervals, these values support practical equivalence rather than a meaningful ranking between bagging and boosting. MLP also retained useful predictive ability (R-squared = 0.7928), whereas the larger ridge-regression error indicates that nonlinear models represented the observed composition-ratio relationships more effectively in this dataset.

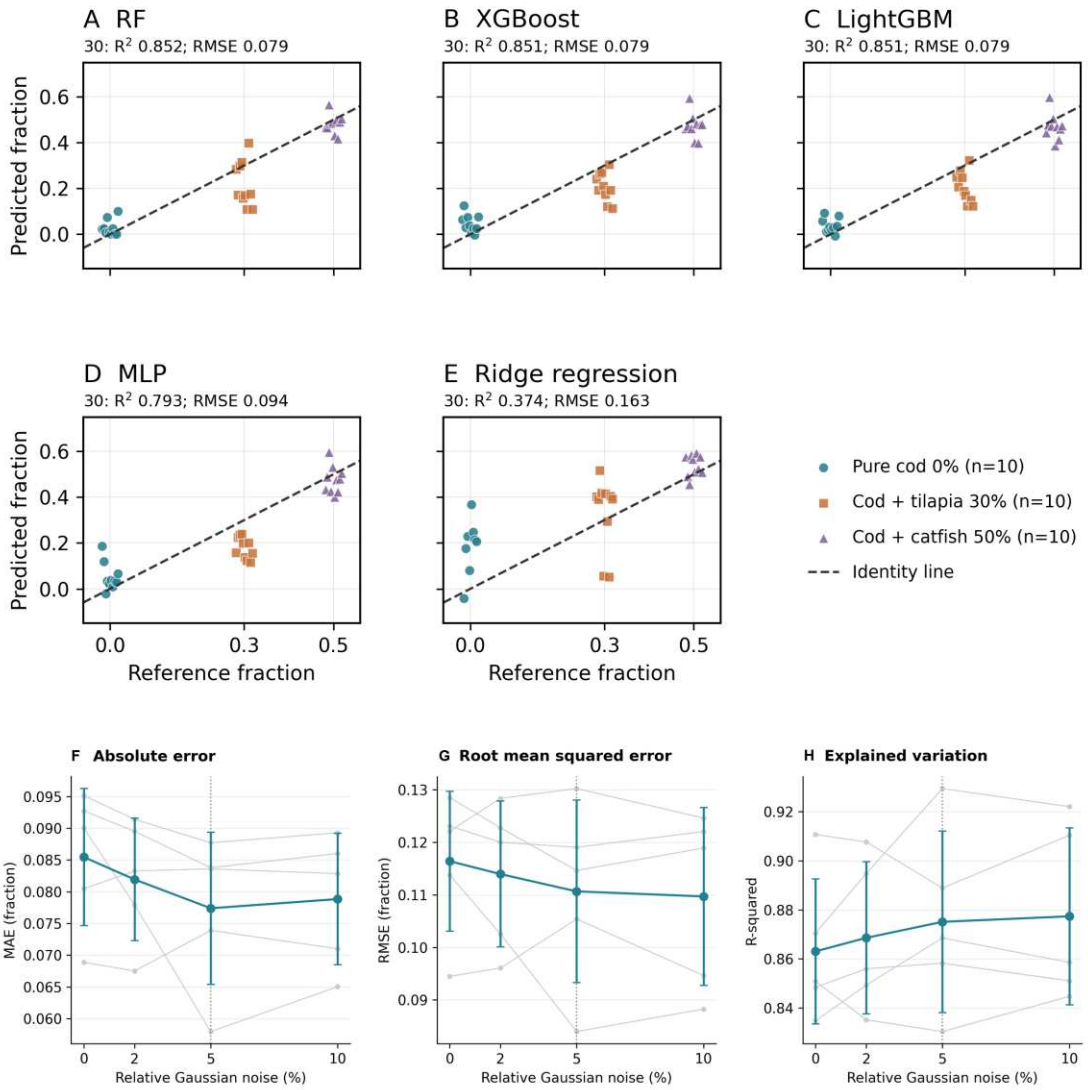
Table 3. Common-regressor performance on the 30 measured validation samples.

Model	MAE	RMSE	R^2
RF	0.0542	0.0791	0.8517
XGBoost	0.0619	0.0792	0.8515
LightGBM	0.0626	0.0794	0.8507
MLP	0.0758	0.0935	0.7928
Ridge regression	0.1312	0.1626	0.3741

All 30 rows are measured samples; this is a validation set. The target is 0 for ten pure-cod samples, 0.30 for ten cod–tilapia mixtures and 0.50 for ten cod–catfish mixtures. Only pooled metrics across all 30 rows are reported. Ridge denotes ridge regression. Predictions are untruncated.

The measured validation deliberately included three formulation groups: ten pure-cod samples, ten cod-tilapia mixtures at 30% adulteration and ten cod-catfish mixtures at 50%. The pooled metrics test whether one frozen model can recover the ratio pattern across these two adulterant species and pure material. Because only three target levels were represented, the pooled R-squared is influenced by separation among the formulation groups and should be interpreted as measured recovery of these defined groups rather than validation of continuous calibration across the full 0-100% interval. Crossed ratio levels for each species and additional independent

528 batches are needed to estimate species-specific calibration, prediction uncertainty and low-level
529 detection performance.



530

531 Figure 7. Measured validation and noise-sensitivity results. Panels A-E show frozen-model predictions
532 for all 30 measured quantitative validation samples: ten pure-cod samples at 0%, ten cod-tilapia mixtures
533 at 30% adulteration and ten cod-catfish mixtures at 50% adulteration. Panels A-E correspond to RF,
534 XGBoost, LightGBM, MLP and ridge regression, respectively. Each panel reports the pooled 30-sample
535 R-squared and RMSE; predictions are untruncated. Panels F-H show the sensitivity of the thirteen-
536 predictor RF regressor to 0%, 2%, 5% and 10% relative Gaussian noise in MAE (F), RMSE (G) and R-
537 squared (H). Training and generated testing use matched noise levels. Error bars represent fold dispersion,
538 not calibrated measurement uncertainty.

The agreement between the internal evaluation and the 30 measured profiles supports the central modelling strategy within the tested formulations. Linear mixing supplied dense virtual coverage from 0% to 100%, while the separately measured samples tested transfer beyond the generated rows at 0%, 30% and 50%. For RF, the MAE of 0.0542 and RMSE of 0.0791 correspond to 5.42 and 7.91 percentage points, respectively, on the adulterant-fraction scale. These errors support further evaluation of the method as a screening approach; whether they are acceptable for a specific control decision depends on a predefined decision threshold, which was not established in this study.

3.7 Noise sensitivity analysis

RF performance varied little across the four assumed noise levels (Table 4; Fig. 7F-H). Mean RMSE changed from 0.1164 at 0% noise to 0.1140, 0.1107 and 0.1097 at 2%, 5% and 10%, respectively, while mean R-squared increased from 0.8631 to 0.8774. The small improvement with increasing noise should not be interpreted as evidence that measurement noise benefits prediction. Each condition applied the same noise assumption to training and simulated testing, so the comparison evaluates sensitivity to matched generator settings rather than robustness to independent laboratory measurement error. Real-world robustness must therefore be assessed with additional measured batches and deliberately varied analytical uncertainty.

Table 4. Noise sensitivity of the 13-predictor RF common regressor.

Noise (%)	MAE	RMSE	R^2
0	0.0854 ± 0.0108	0.1164 ± 0.0133	0.8631 ± 0.0295
2	0.0819 ± 0.0096	0.1140 ± 0.0139	0.8686 ± 0.0311
5	0.0774 ± 0.0120	0.1107 ± 0.0174	0.8751 ± 0.0370
10	0.0788 ± 0.0103	0.1097 ± 0.0169	0.8774 ± 0.0361

Mean ± sample SD across five source folds. Each fold retains its parameters selected at 5% noise. The 5% condition is the main-analysis result, not a separate repeat.

3.8 Model interpretation and limitations

The tree ensembles consistently produced the strongest classification and regression results. RF reduces variance by averaging randomised trees (Breiman, 2001), whereas XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017) learn additive boosted-tree structures. These models can represent thresholds, nonlinear responses and interactions without requiring them to be specified in advance, which is advantageous for a compact tabular dataset containing proximate and elemental variables on different scales. Their closely similar regression

performance indicates that the usable composition signal was not confined to one tree-learning strategy.

The comparison with MLP and the linear baselines further clarifies the structure of the task. MLP captured much of the predictive information, but its greater fold variability and lower measured R-squared suggest that it was less well matched to the limited tabular data. Ridge regression provided a useful linear benchmark, but its larger errors indicate that a single global linear relationship did not represent the pooled cod-tilapia and cod-catfish trajectories adequately. This comparison supports ensemble trees for the present feature set without implying that the same ranking will hold in larger or differently processed datasets.

The virtual-sample design is a key component of the quantitative workflow. Linear mass weighting is a transparent first-order representation for concentration features and provides ratio coverage at 10% intervals, while the 5% perturbation introduces controlled variation around those mixtures. Virtual sample generation can support learning from small datasets (Wen et al., 2025), but generated rows do not create new biological sources and cannot replace independent material. The 30 measured profiles therefore provide the essential transfer test; their R-squared values of approximately 0.85 for all three tree ensembles support the simulation strategy at the three measured formulation levels, while leaving broader calibration to future measured batches.

The composition approach and other seafood-authentication platforms should be viewed as complementary. Spectroscopic and volatile-fingerprint methods can provide rapid or minimally destructive screening (Chen et al., 2024; Feng et al., 2023; Feng et al., 2024; Hou et al., 2019), whereas DNA-based methods are better suited to confirming taxonomic identity in processed or mixed products (Detcharoen et al., 2024; Piredda et al., 2022). Elemental and proximate measurements require laboratory analysis, but they provide interpretable concentration data and can address both source-category classification and mixture-ratio prediction in one modelling framework. The present method therefore complements, rather than replaces, rapid fingerprinting and molecular confirmation.

Several limitations define the scope of these findings. The source library comprised 30 records from three fish categories, and the measured validation contained only 0%, 30% and 50% formulations prepared from a limited set of commercially sourced materials. Species identity relied on commercial labels rather than independent molecular confirmation, and database heterogeneity may reflect differences in origin, season or analytical practice. The pooled measured metrics do not establish a detection limit, a decision threshold, prediction intervals or continuous accuracy across the full ratio range. In addition, processing steps such as washing, heating and additive use may alter the same composition variables used by the models.

Accordingly, the present results support feasibility within the defined cod, tilapia and catfish system rather than immediate generalisation to all commercial surimi. Future validation should

cross adulterant species with additional low and intermediate ratios, independent production lots, seasons and processing conditions; confirm species by molecular methods where possible; and report prediction uncertainty and application-specific decision thresholds. Such studies would determine whether the current composition signal remains stable across supply-chain variation and would permit a formal limit of detection or quantification to be established.

4 CONCLUSIONS

The comparative analysis demonstrates strong performance for physicochemical- and elemental-composition modelling of surimi. RF and LightGBM correctly classified all 30 held-out source records, while XGBoost achieved the lowest internal regression RMSE of 0.1061. The agreement among multiple tree ensembles confirms that the predictive signal is present across model families. Classification used the unaugmented source records, whereas quantitative prediction combined source-partitioned virtual mixtures with a separate measured validation set.

SHAP analysis showed that fat, moisture and zinc contributed prominently to classification, whereas fat, iron and sodium contributed prominently to regression. These patterns agree with published evidence that both lipid-related information and multielement profiles can support fish or surimi differentiation (Han et al., 2021; Liu et al., 2015). The results also demonstrate the benefit of integrating proximate and elemental features rather than using either group alone. Most importantly, direct evaluation of 30 laboratory-measured samples confirmed transfer from the virtual-mixture training framework. RF achieved MAE = 0.0542, RMSE = 0.0791 and R-squared = 0.8517; XGBoost achieved MAE = 0.0619, RMSE = 0.0792 and R-squared = 0.8515; and LightGBM achieved a closely similar R-squared of 0.8507. The combined workflow therefore provides a practical basis for surimi source classification and adulteration-ratio screening. Further measured samples across additional ratios, batches and commercial products can extend the demonstrated application range.

AUTHOR INFORMATION

Corresponding Authors

Qingchao Xie-*Laboratory of Quality & Safety Risk Assessment for Aquatic Product on Storage and Preservation (Shanghai), Ministry of Agriculture and Rural Affairs, College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China*; orcid.org/0000-0002-2926-4304; Email: qcxie@shou.edu.cn

Authors

Linjun Zhou- *College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China*

Lingyun Ma- *College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China*

Xianhai Song - *College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China*

Xinhui Yang - *College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China*

Bian Li - *College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China*

Hongmin Zhang-*College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China*

Yong Zhao - *Laboratory of Quality & Safety Risk Assessment for Aquatic Product on Storage and Preservation (Shanghai), Ministry of Agriculture and Rural Affairs, College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China*

Author Contributions

L.Z.: Methodology, Data curation, Visualization, Formal analysis, Writing—original draft; L.M.: Methodology, Data curation, Visualization; X.S.: Supervision; X.Y.: Methodology, Data curation; B.L.: Writing—review & editing, Sample collection; H.Z.: Funding acquisition, Writing—review & editing; Y.Z.: Writing—review & editing; Q.X.: Investigation, Resources, Supervision, Writing—review & editing.

FUNDING

This study was supported by the National Key Research and Development Program of China (No. 2024YFD2402200), National Natural Science Foundation of China (NO. 32372292).

Declaration of Interest

The authors declare no competing financial interest.

Data Availability Statement

The datasets used and analyzed during the current study are publicly available from the following sources:

1. The United States Department of Agriculture (USDA) Food Data Central database (available at: https://fdc.nal.usda.gov/food-search/).
2. The Food and Agriculture Organization of the United Nations (FAO) FAO/INFOODS uFiSh 1.0 database (available at: https://www.fao.org/food-composition/tables-and-databases/detail/f-food-composition-tables/en).

3. The Food-Nutrition Composition Query Platform published by the National Institute for Nutrition and Health (Chinese Center for Disease Control and Prevention) and the Chinese Nutrition Society (available at: https://nlc.chinanutri.cn/fq/).

Details about the specific database versions, parent food items, food identifiers, and source identifiers for the 30 records used in this study are provided in Supplementary Table S1. Further inquiries regarding the data used in this study can be directed to the corresponding author.

REFERENCES

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. https://www.jmlr.org/papers/v11/cawley10a.html

Charrondiere, U. R., Rittenschober, D., Nowak, V., Stadlmayr, B., Wijesinha-Bettoni, R., & Haytowitz, D. (2016). Improving food composition data quality: Three new FAO/INFOODS guidelines on conversions, data evaluation and food matching. *Food Chemistry*, *193*, 75–81. https://doi.org/10.1016/j.foodchem.2014.11.055

Chen, J., Tang, H., Wang, M., Wei, H., & Ou, C. (2024). Explorative study for the rapid detection of adulterated surimi using gas chromatography-ion mobility spectrometry. *Food Chemistry*, *439*, 138083. https://doi.org/10.1016/j.foodchem.2023.138083

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Detcharoen, M., Khrueakaew, P., Sukkapat, P., Benjakul, S., & Saetang, J. (2024). Metabarcoding for authentication of fish species in surimi-based products by Nanopore sequencing. *Food Bioscience*, *61*, 104628. https://doi.org/10.1016/j.fbio.2024.104628

Feng, S., Qi, S., Chen, C., Khan, I. M., Gu, Q., Zhang, Y., & Wang, Z. (2024). Classification and quantification of adulterants in Alaska pollock based on Fourier transform near-infrared spectroscopy with chemometrics. *Food Bioscience*, *59*, 104003. https://doi.org/10.1016/j.fbio.2024.104003

Feng, Y., Wan, S., Pan, S., Luan, P., Kong, L., & Zhu, M. (2023). Quantitative detection of surimi adulteration based on spectral reconstruction of RGB images. *Transactions of the Chinese Society of Agricultural Engineering*, *39*(20), 275–282. https://doi.org/10.11975/j.issn.1002-6819.202307013

Filonzi, L., Ardenghi, A., Rontani, P. M., Voccia, A., Ferrari, C., Papa, R., Bellin, N., & Nonnis Marzano, F. (2023). Molecular barcoding: A tool to guarantee correct seafood labelling and quality and preserve the conservation of endangered species. *Foods*, *12*(12), 2420. https://doi.org/10.3390/foods12122420

Goyal, R., Singha, P., & Singh, S. K. (2024). Spectroscopic food adulteration detection using machine learning: Current challenges and future prospects. *Trends in Food Science & Technology*, *146*, 104377. https://doi.org/10.1016/j.tifs.2024.104377

Han, C., Dong, S., Li, L., Gao, Q., & Zhou, Y. (2021). Evaluation of elemental analysis assisted by chemometrics for authenticating production methods and geographical origins of salmonids. *Aquaculture*, *545*, 737210. https://doi.org/10.1016/j.aquaculture.2021.737210

Han, C., Li, L., Dong, X., Gao, Q., & Dong, S. (2022). Current progress in the authentication of fishery and aquatic products using multi-element and stable isotope analyses combined with chemometrics. *Reviews in Aquaculture*, *14*(4), 2023–2037. https://doi.org/10.1111/raq.12686

Hou, S.-W., Wei, W., Wang, Y., Gan, J.-H., Lu, Y., Tao, N.-P., Wang, X.-C., Liu, Y., & Xu, C.-H. (2019). Integrated recognition and quantitative detection of starch in surimi by infrared spectroscopy and spectroscopic imaging. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *215*, 1–8. https://doi.org/10.1016/j.saa.2019.02.080

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. *Advances in Neural Information Processing Systems*, *30*, 3146–3154. https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Liu, Y., Hu, W., Guo, X.-X., Wang, X.-C., Sun, S.-Q., & Xu, C.-H. (2015). Rapid discrimination of three marine fish surimi by Tri-step infrared spectroscopy combined with principle component regression. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *149*, 516–522. https://doi.org/10.1016/j.saa.2015.04.116

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Piredda, R., Mottola, A., Cipriano, G., Carlucci, R., Ciccarese, G., & Di Pinto, A. (2022). Next generation sequencing (NGS) approach applied to species identification in mixed processed seafood products. *Food Control*, *133*, 108590. https://doi.org/10.1016/j.foodcont.2021.108590

Ren, L., Tian, Y., Yang, X., Wang, Q., Wang, L., Geng, X., Wang, K., Du, Z., Li, Y., & Lin, H. (2023). Rapid identification of fish species by laser-induced breakdown spectroscopy and Raman spectroscopy coupled with machine learning methods. *Food Chemistry*, *400*, 134043. https://doi.org/10.1016/j.foodchem.2022.134043

Sampels, S. (2015). The effects of processing technologies and preparation on the final quality of fish products. *Trends in Food Science & Technology*, *44*(2), 131–146. https://doi.org/10.1016/j.tifs.2015.04.003

Sueker, M., Daghighi, A., Akhbardeh, A., MacKinnon, N., Bearman, G., Baek, I., Hwang, C., Qin, J., Tabb, A. M., Roungchun, J. B., Hellberg, R. S., Vasefi, F., Kim, M., Tavakolian, K., & Kashani Zadeh, H. (2023). A novel machine-learning framework based on a hierarchy of dispute models for the identification of fish species using multi-mode spectroscopy. *Sensors*, *23*(22), 9062. https://doi.org/10.3390/s23229062

Varrà, M. O., Ghidini, S., Husáková, L., Ianieri, A., & Zanardi, E. (2021). Advances in troubleshooting fish and seafood authentication by inorganic elemental composition. *Foods*, *10*(2), 270. https://doi.org/10.3390/foods10020270

Wen, J., Su, A., Wang, X., Xu, H., Ma, J., Chen, K., Ge, X., Xu, Z., & Lv, Z. (2025). Virtual sample generation for small sample learning: A survey, recent developments and future prospects. *Neurocomputing*, *615*, 128934. https://doi.org/10.1016/j.neucom.2024.128934

Zaldarriaga Heredia, J., Wagner, M., Jofré, F. C., Savio, M., Azcarate, S. M., & Camiña, J. M. (2023). An overview on multi-elemental profile integrated with chemometrics for food quality assessment: Toward new challenges. *Critical Reviews in Food Science and Nutrition*, *63*(26), 8173–8193. https://doi.org/10.1080/10408398.2022.2055527

Zhang, X., Giusti, A., Sun, Z., Li, Y., Guo, J., Deng, W., Chen, Y., He, A., Peng, H., Tinacci, L., Armani, A., & Wen, J. (2024). Molecular authentication of surimi-based products (fish cakes, 鱼糕) sold on the Chinese e-commerce: Traditional (DNA barcoding) and innovative techniques (metabarcoding) to tackle seafood fraud. *Food Control*, *155*, 110110. https://doi.org/10.1016/j.foodcont.2023.110110

Zhang, Z., Li, Y., Zhao, S., Qie, M., Bai, L., Gao, Z., Liang, K., & Zhao, Y. (2024). Rapid analysis technologies with chemometrics for food authenticity field: A review. *Current Research in Food Science*, *8*, 100676. https://doi.org/10.1016/j.crfs.2024.100676

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS11.docx