Vol. 4, No. 2, 2026
DOI: 10.61975/gjset.v4i2.142

REVIEW

Preprocessing, Feature Selection, and Validation Strategies for FTIR and GC-MS Chemometric Models in Food Authentication: A Structured Critical Review

Taufiq Karma[1], Rinaldi Indroes[2,*], Khairan Khairan[2], and Mohamad Rafi[3]

[1]Graduate School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; [2]Department of Pharmacy, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; [3]Department of Chemistry, Faculty of Mathematics and Natural Sciences, IPB University, Bogor 16680, Indonesia

* Correspondence: rinaldi.idroes@usk.ac.id

Article History

Received
24 August 2026

Accepted
13 October 2026

Available Online
30 October 2026

Keywords

Food authentication
FTIR
GC-MS
Chemometrics
Preprocessing

Abstract

Fourier transform infrared (FTIR) spectroscopy and gas chromatography–mass spectrometry (GC-MS), combined with chemometrics and machine learning, are increasingly applied to food authentication. However, model reliability and generalizability depend on preprocessing, feature selection, and validation strategies, particularly when handling high-dimensional analytical data susceptible to overfitting and information leakage. This systematic critical review examined these strategies in FTIR- and GC-MS-based food-authentication models. A systematic search of ScienceDirect was conducted for publications from 2016 to 2026 following PRISMA 2020 guidelines. Of 274 identified records, 74 studies met the eligibility criteria, comprising 45 FTIR-based studies (60.8%) and 29 GC-MS-based studies (39.2%). Common FTIR preprocessing methods included derivatives, standard normal variate, Savitzky–Golay smoothing, mean-centering, and scatter correction, whereas GC-MS studies more frequently employed scaling, normalization, alignment, and peak processing. Feature-selection approaches varied considerably, with variable importance in projection (VIP), loading-based methods, and recursive feature elimination among the most frequently reported. External or independent validation was reported in 21 studies (28.4%). Explicit information-leakage prevention was documented in only three studies (4.1%), while six (8.1%) raised potential leakage concerns and 65 (87.8%) provided insufficient or unclear reporting. Tree-based algorithms were used in 31 studies, predominantly Random Forest, whereas standalone Decision Tree models were less frequently reported. Overall, no preprocessing or feature-selection strategy demonstrated universal superiority across analytical techniques and authentication tasks. Reliable food-authentication models require an integrated, leakage-aware workflow in which data partitioning precedes preprocessing, feature selection, and data-driven model optimization, followed by evaluation using independent samples whenever possible. Future research should prioritize methodological transparency, reproducibility, interpretability, robustness, and external generalizability rather than focusing primarily on internal predictive performance.

Introduction

Food authentication is becoming increasingly important as food fraud takes place in an increasingly complex global supply chain, driven by economic motives, and can harm consumers, industry, and regulators [1–3]. Adulteration, substitution, mislabeling, and dilution most often target high-value commodities such as milk, meat, vegetable oil, honey, coffee, and powdered products, requiring rapid, sensitive, and routinely applicable detection methods [3,4].

In this context, Fourier transform infrared (FTIR) and gas chromatography–mass spectrometry (GC–MS) are two very interesting platforms because they both produce information-rich chemical fingerprints to distinguish between authenticity, geographical origin, production

Copyright: © 2026 by the authors. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License. (https://creativecommons.org/licenses/by-nc/4.0/)

process, and the existence of adulterants [5–7]. FTIR is superior as a fast, non-destructive, relatively inexpensive, and easily integrated method into quality control, while GC-MS offers deeper sensitivity, specificity, and coverage of volatile and semi-volatile metabolites [5,8].

The power of FTIR and GC-MS actually creates a huge analytical challenge because both of these techniques produce very large amounts of multivariate data, both at the level of the number of variables per sample and the number of samples, batches, and acquisition conditions that the model must cover [9–11]. In spectroscopy, a single spectrum often contains more than a thousand wavelengths, whereas fingerprinting and untargeted chromatography approaches can produce thousands of chemical features from a single sample [12–14].

In FTIR, complexity arises due to overlapping absorption bands, baseline drift, scattering effects, heterogeneity of composition, and nonlinear variability, especially when the differences between classes are very subtle [15,16]. In GC-MS and related chromatography platforms, complexity comes from the not always perfect peak resolution, the possibility of intercompound overlap, the need for alignment, as well as the high amount of signals generated by non-targeted and multidimensional chromatography approaches [11,13,17].

The core problem with high-dimensional data is not only the number of variables, but also multilinearity, information redundancy, and the imbalance in the ratio of features to sample numbers [18,19]. The NIR and MIR spectra are known to have high proportions of correlated variables, so approaches such as PLS are often effective, but not always sufficient when the chemical matrix is very complex or intervariable correlation is weak [19].

When the number of features far exceeds the number of samples, authentication models are very vulnerable to overfitting, which captures noise or random structures instead of stable authentication signals [13]. This risk is increasing in food authentication studies that often use tens to hundreds of samples, but model thousands of spectral or chromatographic variables in a single classification pipeline [14,20,21].

The limitations of classical linear projection methods also began to be seen when data were loaded with high noise, local patterns, or nonlinear structures that were not easily captured by global representations based on variance [15]. Therefore, the main challenge is not just to build models with high accuracy, but to extract really relevant, stable, and interpretable information from complex data [22].

In much of the literature, preprocessing is seen as an essential stage because raw spectral signals are easily influenced by environmental noise, instrument artifacts, sample impurities, scattering effects, and radiation distortion that can shift feature extraction in the wrong direction [16,23]. Baseline corrections, normalization, smoothing, derivatives, SNV, MSC, OSC, and their derivatives are used to suppress physical artifacts and highlight relevant chemical variations, but the impact is highly dependent on the dataset, and there is no one universal strategy [24,25].

Comparative evidence suggests that preprocessing can substantially alter model performance; in the evaluation of 108 preprocessing schemes in 26 NIR problems, the correct scheme can lower the RMSEP by 50-60% compared to the raw spectrum, but the heterogeneity between datasets affirms that preprocessing selection should be validated, not assumed [24]. Even if there are specific applications, the best strategies can be different: MSC improves the prediction of manga quality [26], local SNV improves linearity and interpretation [23], whereas fusion scatter correction sometimes outperforms a single correction [27].

Feature selection becomes equally important because dimensional reduction is not only a matter of computational efficiency, but also a way to limit the complexity of the model, suppress overfitting, and find chemically meaningful markers [13]. On LC-MS non-targeted honey data, recursive feature elimination reduced 2028 features to 54 without losing classification

performance, indicating that most initial variables were redundant or uninformative for authentication tasks [14].

The main challenges of FTIR and GC-MS high-dimensional data in food authentication can be summarized into five issues: the number of interrelated variables, the high noise and artifacts of the instrument, the overlap of signals and heterogeneity of the food matrix, the imbalance between the number of features and samples, as well as the weak model transferability when validation is unrealistic. Therefore, preprocessing, feature selection, and validation strategies are not an additional stage, but a methodological foundation that determines the reliability of the chemometric model for food authentication.

Although PLS-DA, SVM, and Random Forest are more commonly used in FTIR and GC-MS-based classifications, in this context, tree-based modeling, which includes Decision Tree as well as ensemble methods such as Random Forest, bagging, and boosting, is one of a group of relevant approaches to study in instrumental data-based food authentication. The Decision Tree provides a conceptual rejection point that is relatively easy to interpret because classification decisions can be traced through the rules of separation of variables, while the ensemble method extends the principle to increase stability and predictive ability. However, the performance of the tree-based model remains influenced by data representation, feature dimensionality, preprocessing strategy, variable selection, and validation design. Therefore, tree-based modelling provides relevant context to evaluate the linkages between these stages in the development of a robust and generalizable model of food authentication.

A number of previous reviews have made important contributions to the development of chemometric analysis and spectroscopy-based food authentication techniques [28]. In particular review variable selection strategies in the chemometric analysis of food data, whereas Rodionova et al. [22] discusses a chemometric approach to food authentication with a focus on the development of reliable models. Yan [16] provides a comprehensive review of spectral data preprocessing for machine learning and quantitative analysis, including baseline correction, scattering correction, normalization, filtering and smoothing, spectral derivatives, as well as advanced preprocessing methods. Smolinska et al. [29] focus on a chemometric strategy for the preprocessing and automation of GC-MS data analysis, while Vicente et al. [5] expand the study of NMR, FTIR, and NIR-based agro-food authentication by discussing preprocessing, modeling, validation, robustness, transferability, and data fusion.

Although previous reviews have discussed spectroscopy, chromatography, chemometrics, and machine learning in food authentication, few have examined preprocessing, feature selection, and validation as interconnected steps in FTIR- and GC-MS-based modeling. These decisions can affect model performance, information leakage, overfitting, and generalizability. Their implications for tree-based methods, including Decision Tree, Random Forest, bagging, and boosting, also remain insufficiently synthesized in food authentication. Therefore, an integrated review is needed to assess not only internal predictive performance but also the reliability of the overall analytical workflow.

This review focuses on FTIR and GC-MS applications in food authentication and adulteration detection, particularly in milk, meat, vegetable oils, honey, and coffee. Relevant methodological studies from non-food applications are also considered as supporting evidence for preprocessing, feature selection, dimensionality reduction, modeling, and validation, but are not treated as primary evidence of food authentication. Accordingly, this review critically examines preprocessing, feature selection, and validation strategies in FTIR- and GC-MS-based chemometric models, focusing on their implications for model robustness, interpretability, information leakage, and generalizability. Particular attention is given to tree-based methods, with Decision Tree considered for its interpretability and ensemble methods examined as

alternative modeling approaches, without assuming that tree-based models dominate the literature.

Materials and Methods

Review Design and Reporting Framework

This review critically synthesizes methodological practices reported in food-authentication studies using FTIR spectroscopy and/or GC–MS with chemometric or machine-learning analysis. Study identification, screening, and reporting were organized with reference to PRISMA 2020, while search reporting was informed by PRISMA-S [30,31]. These reporting guidelines were not used as instruments for evaluating the methodological quality of individual studies. The review was not prospectively registered, and no publicly deposited a priori protocol was available. Consequently, the review is presented as a structured, reproducible synthesis of a defined ScienceDirect-derived corpus rather than a comprehensive estimate of methodological practices across all publication databases.

Reviewing Questions

RQ1. Which preprocessing approaches are reported in FTIR- and GC–MS-based food-authentication studies, and what within-study evidence is available regarding their influence on model performance and interpretation?

RQ2. What feature-selection and variable-selection strategies are used in FTIR- and GC-MS-based chemometric modeling, and how do these strategies relate to model complexity and performance?

RQ3. How do validation strategies influence the assessment of model robustness and generalizability in FTIR- and GC-MS-based food authentication, including studies using tree-based methods?

Information Source and Search Scope

The study corpus was identified through ScienceDirect searches conducted on 9 September 2026, targeting publications dated 2016–2026. Searches used the platform's "Title, abstract, or author-specified keywords" field. Six recorded search strategies addressed FTIR, GC–MS, chemometrics, model validation, and tree-based methods; their exact search strings and record counts are reported in Table 1 and the search audit files. ScienceDirect was the sole source used for the systematic search. The strategy included outcome- or model-related terms, such as "validation" and "Random Forest," which may have favored studies explicitly reporting these features in their searchable metadata. Relevant studies indexed in other databases, or studies that did not include the specified terms in their titles, abstracts, or author keywords, may therefore have been missed. Accordingly, the reported frequencies describe only the included ScienceDirect-derived corpus and should not be interpreted as prevalence estimates for the wider food-authentication literature. Publications identified outside the systematic search were used only for background and interpretation and were not counted among the 74 included studies.

Eligibility Criteria and Boundary Cases

Eligible reports were English-language original research articles dated 2016–2026 that investigated a food or food-relevant matrix, addressed authentication, adulteration, origin, variety/species discrimination, or traceability, and applied multivariate analysis, chemometrics, or machine learning to FTIR and/or GC–MS data. Studies using exclusively non-FTIR/non-GC–MS analytical inputs, lacking an authentication-related objective, or reporting only descriptive chemical characterization were outside scope; reviews and other non-original publication types were excluded. Studies combining analytical platforms require an explicit determination of

whether FTIR or GC–MS contributed to the model under review, rather than serving solely as an unrelated reference measurement. Essential-oil and other borderline matrices require consistent case-by-case justification of their food-authentication relevance. The inclusion status of such borderline studies is subject to the eligibility audit; no retrospective re-screening is claimed here.

To ensure consistent eligibility decisions, studies were required to address an explicit food-authentication objective involving a food, food ingredient, or food-related material with a clearly established connection to food authenticity. Potential use in the food industry alone was not sufficient for inclusion. Studies on materials with multiple industrial applications were eligible only when the investigated application and authentication objective were explicitly relevant to food authenticity. Studies addressing non-food applications without a clear food-authentication objective were excluded.

Record Management and Selection

Search records were exported in RIS format and consolidated in a master database. Duplicate identification used normalized DOI information and, where necessary, normalized titles. Title/abstract screening was followed by full-text eligibility assessment. The existing study-selection workflow involved four reviewers, with uncertain cases discussed to reach consensus. This description does not imply that all records were independently screened in duplicate or that inter-rater agreement was calculated. The recorded flow was 274 identified records, 59 duplicates removed, 215 screened, 100 excluded at title/abstract, 115 reports sought, 30 reports not retrieved, 85 full texts assessed, 11 full texts excluded, and 74 studies included. Reasons for exclusion should match the record-level audit and PRISMA figure exactly. The 30 reports not retrieved represent a substantial potential source of retrieval bias. Of the 115 reports sought for retrieval, 30 could not be obtained for full-text assessment. Therefore, their eligibility and methodological characteristics could not be verified, and they were not included in the final synthesis.

Data Extraction and Classification

A structured extraction database captured study identifiers, food matrix, authentication objective, analytical platforms, sample design, preprocessing, feature-related methods, model types, validation design, performance metrics, and information relevant to leakage and generalizability. FTIR and GC–MS roles should be distinguished from auxiliary analytical platforms. Feature selection producing a reduced, refitted model should be distinguished from variable interpretation (e.g., VIP or loading scores) and feature extraction (e.g., PCA scores or learned representations). Model use should be distinguished from using a model solely to rank variables. Unresolved entries in the extraction database are retained as requiring verification; absence of an extractable detail must not be equated with non-reporting in the primary article unless the full text has been checked.

Studies were classified into two categories: FTIR-based and GC–MS-based studies. Classification was determined by the analytical platform used to generate the data for the chemometric authentication model evaluated in each study. Studies using FTIR data for model development or evaluation were classified as FTIR-based, whereas studies using GC–MS data for these purposes were classified as GC–MS-based. Additional analytical techniques used for preliminary analysis, comparison, or complementary characterization did not automatically determine the classification. Where the primary platform could not be established from the available information, the classification was marked as unverified pending confirmation.

Validation, Leakage, and Units of Independence

Internal hold-out testing was distinguished from external validation. A random split of observations originating from the same study does not by itself establish independent external

validation, even when termed an “external test set” by the original authors. External validation requires a substantively independent source, acquisition period, site, batch, or comparable design whose independence is documented. Where reported, the separation unit (biological sample, technical spectrum/injection, batch, or site) should be extracted, because placing technical replicates of the same underlying sample in both training and test partitions may inflate apparent performance. Preprocessing operations performed separately on each spectrum (such as spectrum-wise SNV or Savitzky–Golay filtering) should be distinguished from dataset-dependent operations (such as fitted centering/scaling, MSC reference estimation, PCA, imputation, supervised selection, or resampling). Leakage control is classified as explicitly addressed only where the study reports an appropriate separation or fitting strategy for the relevant dataset-dependent steps. Otherwise, the evidence is coded as unclear; “unclear” does not establish that leakage occurred. A “potential concern” designation requires an identifiable mechanism by which test-fold information could influence model development.

Methodological-Quality Appraisal

The supplementary workbook contains a reviewer-developed appraisal covering ten domains: sampling representativeness; sample size and class balance; preprocessing reporting and sequence; feature-related method and timing; leakage control; validation design; independent external validation; performance metrics; reproducibility; and proportionality of generalization claims. Each domain is scored 0–2 (total 0–20); the working categories are high (15–20), moderate (10–14), and low (0–9). This rubric is not a validated risk-of-bias tool, and the cutoffs are operational rather than empirically calibrated. The availability of completed scores in the updated supplementary workbook does not establish that two assessors independently rated every study. The scoring process, evidence supporting individual ratings, and any sensitivity analysis must be confirmed before the corresponding results are treated as verified.

Evidence Synthesis and Interpretation

Descriptive frequencies summarize reported practices within the included corpus, with denominators made explicit and modality-specific denominators used where appropriate. For RQ1 and RQ2, evidence about performance is interpreted through within-study comparisons only where the original report provides comparable metrics, datasets, and validation conditions; heterogeneous accuracy, RMSE, and classification metrics are not pooled without methodological justification. Validation categories and tree-based model use are described separately from their potential inclusion as search terms. Differences between internal and external performance are discussed only when both estimates are available for a comparable task. No causal effect of preprocessing or feature selection is inferred from across-study frequency counts.

Results and Discussion

Selection of Studies and Characteristics of The Studies Covered

The search identified 274 records. After removal of 59 duplicates, 215 unique records underwent title and abstract screening. A total of 100 records were excluded at this stage, and 115 reports were sought for full-text retrieval. Thirty reports could not be retrieved, leaving 85 full-text articles for eligibility assessment. A total of 11 full-text articles were excluded, yielding 74 studies for inclusion in the final review. Reasons for full-text exclusion were an ineligible domain/matrix (n = 4), an ineligible analytical technique (n = 3), an ineligible publication type (n = 2), and an ineligible application/research objective (n = 2).

The included studies addressed adulteration detection/quantification, discrimination of geographical and botanical origin, species or variety identification, authenticity verification, and traceability. Following reassessment of the five studies previously grouped as multimodal/other,

45 studies (60.8%) were classified as FTIR-based and 29 (39.2%) as GC-MS-based. Studies were classified according to the FTIR- or GC-MS-based authentication model evaluated within the study; additional analytical platforms were not treated as a separate category. Based on the current methodological assessment, 16 studies (21.6%) were classified as high quality, 49 (66.2%) as moderate quality, and 9 (12.2%) as low quality. Detailed study characteristics are presented in Tables 1 and 2.

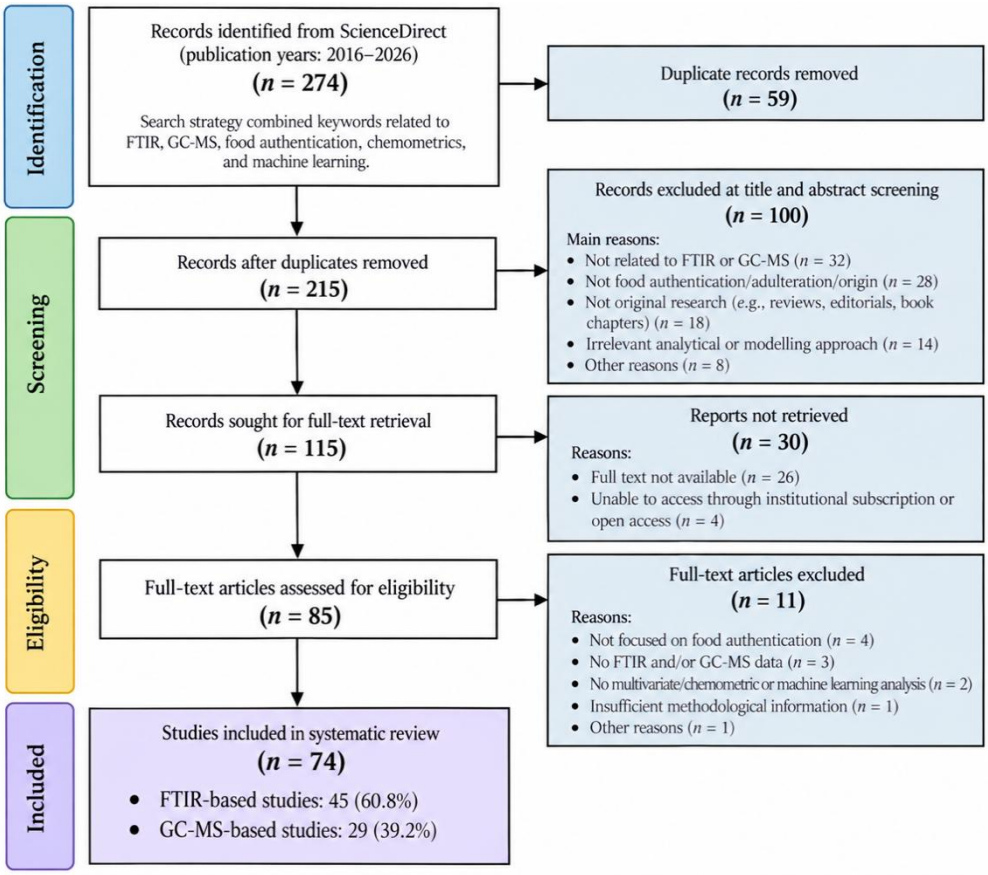


Figure 1. PRISMA flow diagram.

Table 1. Characteristics of FTIR-based food-authentication studies included in the review (n = 45).

No.	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML Models	Validation Strategy	Ref.
1	Milk powder	Adulteration detection	FTIR plus Raman / fused spectroscopy	FTIR: SG+MSC+2D reported best; Raman: SG+MSC+SNV	CARS; mutual information; LLDF/fusion schemes	Multiple ML classifiers/regressors	5-fold cross-validation; test-set evaluation	[32]
2	Raw milk adulterated with cheese whey	Adulteration detection/quantification	FTIR	Spectral preprocessing before modelling; exact variants not fully specified in the extracted methods	Not clearly isolated as dedicated feature selection in current extraction	CART; ANN and chemometric models	Training/validation strategy reported; exact resampling details not fully specified in the extracted methods	[33]
3	Turmeric powder adulterated with Sudan I	Adulteration detection	FTIR	FTIR spectral preprocessing to improve quality/reduce noise	Model-dependent/random feature selection noted for RF; dedicated spectral FS not central	ANN; SVM; RF; KNN	Grid search + cross-validation; separately prepared external validation	[34]
4	Coriander powder + sawdust	Adulteration quantification	ATR-FTIR	SG smoothing; SNV; EMSC; raw data compared	RFE; PCA feature extraction	LR; DT; SVR; ANN	10-fold CV with 80/20 split; validation and testing metrics	[35]
5	Ground coffee + barley/chickpea/date pit	Adulteration detection/quantification	FTIR	StandardScaler; SNV; MSC; SG smoothing; 1st derivative; 2nd derivative; wavelet; median filter	No dedicated feature selection central to study	Bagging; RF; Gradient Boosting; XGBoost	70/30 train-test; model optimization	[36]
6	Barrantes wine / Folla Redonda	Wine adulteration detection and quantification	FTIR plus Vis-NIRS	SNV; normalization; derivative transformations and other spectral pretreatments	Recursive Feature Elimination (RFE)	SVM; PLS-DA; RF; regression models	5-fold cross-validation; classification/regression evaluation	[37]
7	Milk	Authentication of thermal-processing intensity / label claims	FTIR	Mean centering; 1st-order SG derivative; 2nd-order SG derivative; 1st SG selected	PCA-loading-based wavenumber selection	kNN; SVM; RF; LDA	60% train / 20% validation / 20% test; grid-search CV; repeated stratified 10-fold CV for statistical comparison	[38]
8	Soy-based beverages + added water	Adulteration detection	ATR-FTIR	Mean centering; SNV; first derivative and combinations; model-specific pretreatment	No separate spectral feature-selection step central; RF performs random	UNEQ; SIMCA; DD-SIMCA; OC-RF; kNN; PLS-DA; DC-RF; DC-RF-MC	Kennard-Stone train/test for some models; 5-fold CV; Monte-Carlo 80/20 train-test for RF	[39]

No.	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML Models	Validation Strategy	Ref.
					variable selection internally			
9	Mediterranean honey	Geographical and botanical origin authentication	ATR-FTIR	First derivative; mean centering; spectral range 1800–750 cm^{-1}	Most relevant wavelengths selected based on cross-validation accuracy	PCA; RF; DD-SIMCA	80/20 train-test; 10-fold CV	[40]
10	Sausage / meat species	Species authentication and mislabeling detection	FTIR-ATR	EMSC normalization; scaling; PCA performed within training folds	PCA dimensionality reduction	Logistic Regression; KNN; SVM; RF	70/30 split; 5-fold CV only on training data	[41]
11	Carob honey adulterated with glucose/rice syrup	Adulterant-type classification and concentration quantification	ATR/mid-FTIR	Five protocols including derivatives, smoothing and baseline correction	No dedicated feature-selection method central to the study	LogR; RFC; SVC; LR; RFR; SVR	10-fold CV with distinct training/validation/blind-test strategy	[42]
12	Honey adulterated with glucose/rice syrup	Adulteration classification	ATR-FTIR	SG second derivative (window 15, polynomial 3) + StandardScaler	RF feature importance identifies discriminative spectral regions	KNN; DT; RF; XGBoost; Gradient Boosting	Biological-sample-level 80/20 split; stratified 5-fold CV on training set	[43]
13	Evening primrose oil (plus NIR nutmeg case)	Adulteration detection using one-class classification	ATR-FTIR for primrose oil; NIR for nutmeg	Infrared spectral preprocessing used in one-class workflow; exact FTIR pretreatment retained for final QC	Artificial outlier generation; RF variable subsampling	One-class RF; PLS-DA; SIMCA	External validation with adulterated and different oil types	[44]
14	Chia oil	Adulteration detection/quantification and α-linolenic acid prediction	FTIR-ATR	First-order polynomial baseline correction for PCA; mean-centering for optimized PLS models	PLS latent-variable modelling; VIP score interpretation	PCA; PLS regression	KS/SPXY calibration-prediction partition + 10-fold CV for latent-variable selection	[45]
15	Buffalo milk adulterated with cow milk	Adulteration detection	FTIR-ATR	FTIR spectral processing; reduced measurement/features evaluated	Particle Swarm Optimization (PSO) for feature reduction	SIMCA; DD-SIMCA; six AI algorithms including Ensemble Bagged Tree	Model training/testing reported; exact split/CV details final QC	[46]
16	Milk and ripened cheese	Farming-system-of-origin authentication	FTIR spectra for milk; other platforms for cheese	FTIR spectral handling prior to discriminant analysis; exact pretreatment final QC	No dedicated feature selection central to main FTIR analysis	LDA and multi-source data combination	10-fold cross-validated LDA	[47]

No.	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML Models	Validation Strategy	Ref.
17	Mentha arvensis essential oil	Adulteration/authentication of high-value essential oil used in food and other industries	ATR-FTIR	Multiplicative Scatter Correction (MSC)	PCA dimensionality reduction; no dedicated variable-selection algorithm reported	PCA; PCA-LDA; PLS-DA; SIMCA; PLSR; PCR	Classification/regression validation reported; exact split final QC	[48]
18	Six food-related FTIR benchmark datasets	Food authentication / classification benchmark	FTIR benchmark datasets	No additional dataset-specific preprocessing	Sliding-window embeddings + topological data analysis multi-scale feature extraction	TDA descriptors + bagging ensemble; comparison with 20 classifiers	Official fixed train-test splits	[15]
19	Black tea	Geographical-origin authentication / fraud screening	FTIR + NIR + XRF with data fusion	Standardization/scaling per platform; information-, feature-, and decision-level fusion	RFE and VIP-RFE; targeted variable selection	LDA; SVM; RF; 1D-CNN; MLP; PLS-DA	30% independent test; remaining data split calibration/validation; 7-fold CV for PLS-DA	[49]
20	Green Arabica and Robusta coffee beans	Species discrimination / adulteration screening	ATR-FTIR	SG smoothing; MSC; SNV; SG performed best	PCA dimensionality reduction; no dedicated feature-selection algorithm	PCA; PCA-LDA	Classification evaluation reported; exact external split not prominent	[50]
21	Coconut water	Sugar substitution/adulteration detection	ATR-FTIR	Savitzky–Golay smoothing + gap-segment derivatives; fingerprint 1200–950 cm^{-1}	PLS latent-variable reduction; no dedicated feature-selection algorithm	PCA; PLS regression	80/20 train-test split; LOOCV for PLS	[51]
22	Moroccan honey	Artificial-honey adulteration authentication and quantification	ATR-FTIR	GLSW preprocessing; comparison with unprocessed spectra	No dedicated feature-selection method central	PCA; PLS-DA; PLSR	Calibration/validation modelling reported	[52]
23	Philippine honey	Geographical and entomological-origin discrimination	FTIR	No additional spectral preprocessing	No dedicated feature selection	PCA; HCA	Exploratory unsupervised discrimination; no predictive validation	[53]
24	Sesame oil adulterated with vegetable oils	Adulteration-level prediction	FTIR	Systematic comparison of OSC, SNV, EMSC and combinations; detrending etc.	No dedicated variable-selection method	PLSR	Calibration/prediction modelling; preprocessing strategies compared	[54]

No.	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML Models	Validation Strategy	Ref.
25	Red chilli powder	Sudan IV adulteration detection	ATR-FTIR	Savitzky–Golay filtering; polynomial/window combinations evaluated	PC-number/model-complexity selection; no separate wavelength-selection algorithm	DD-SIMCA one-class classification	One-class model validation across preprocessing settings	[55]
26	Saffron	Chemical-adulterant/purity quantification	DRIFTS/FTIR	Moving-average smoothing + SNV	ACE transformations; no conventional wavelength-selection algorithm central	ACE; PLS-R; SVM-R	Duplex calibration/prediction split	[56]
27	Chinese yam (Dioscorea spp.)	Species authentication/adulterant-species discrimination	ATR-FTIR	ATR correction; smoothing; baseline correction; normalization; arithmetic processing	Latent-variable reduction through PCA/OPLS-DA; CNN learns features automatically	PCA; PCA-Class; OPLS-DA; CNN	Cross-validation/prediction metrics plus random segregation test	[57]
28	Crude palm oil	Geographical and temporal-origin traceability	FTIR	SNV + Pareto scaling selected as optimal	OPLS-DA loadings identify discriminating spectral regions	OPLS-DA	Blind-sample prediction models	[58]
29	Edible vegetable oils	Processing-technique discrimination and rapeseed-oil adulteration detection	FTIR + Raman	Minimal/limited preprocessing; fusion approach compared with FTIR-only models	1D-CNN automated feature extraction; attention mechanism	LDA; SVC; multi-input 1D-CNN	Train/test evaluation	[59]
30	Honey	Syrup-adulteration classification and quantification	FTIR-HATR	Spectral preprocessing in chemometric pipeline	Feature selection explicitly evaluated; PLS-based feature selection for regression	GBDA; SVMDA; GBR; other classification/regression algorithms	External validation plus model tuning/evaluation	[60]
31	Blackcurrant, strawberry and raspberry seed oils	Commercial authenticity screening	FTIR + DSC	Personalised FTIR preprocessing; selected	VIP selection explicitly compared	PLS-DA; SIMCA	Cross-validation plus external validation with commercial oils	[61]

No.	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML Models	Validation Strategy	Ref.
				spectral regions; noise/baseline treatment	before/after PLS-DA			
32	Edible cricket, mealworm and buffalo-worm flours	Species authenticity/adulteration screening	FTIR plus two portable NIR instruments	SNV and 1st Savitzky–Golay derivative tested individually and combined; mean centering	PCA latent variables; no separate wavelength-selection algorithm central	PCA; DD-SIMCA; soft PLS-DA; PLSR	Training/test evaluation	[62]
33	Polish honey	Botanical/varietal differentiation	FTIR-ATR	FTIR spectral processing for PCA/LDA; exact nesting not explicit	PCA dimensional reduction; no dedicated variable-selection method	PCA; LDA	LOOCV	[63]
34	UHT liquid milk	Hydrogen-peroxide adulteration detection and quantification	ATR-FTIR	Comparison of full spectrum and selected wavenumber windows	Spectral-window selection; fingerprint around 2826 cm^{-1} identified	PCA; DA; MLR	Cross-validation plus testing/external validation	[64]
35	Virgin coconut oil	Palm-oil adulteration detection/quantification	ATR-FTIR	Normal, 1st and 2nd derivatives; spectral-region comparison	Spectral-region selection through informative domains	LDA; PCR; PLS-R	Cross-validation plus prediction evaluation	[65]
36	Coconut milk	Water-adulteration detection	FTIR/mid-infrared	Five spectral treatment techniques compared	Latent-variable reduction in PCR/PLSR; no dedicated feature selector central	PCR; PLSR; SVMR	LOOCV for calibration then predictive-data testing	[66]
37	Edible oils / Camellia oil blends	Oil identity and adulteration authentication	FTIR	Spectral-domain processing; optimal informative intervals	Backward interval PLS (BiPLS) selected 3100–2900, 1800–1700, 1500–1400 and 1200–1100 cm^{-1}	Discriminant analysis; BiPLS quantitative model	External validation	[67]
38	Cashew nuts	Detection of up to four nut adulterants	ATR-FTIR plus NIR	Multiple signal-processing methods tested sequentially	ROC-based class-limit optimization; high-level data fusion	One-class SIMCA	Dedicated training/test sets; ROC optimization on test distances	[68]

No.	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML Models	Validation Strategy	Ref.
39	Meat	Fresh versus frozen/thawed authentication (food fraud)	FTIR plus multispectral imaging	Spectral processing followed by PLS-based variable transformation/selection	PLS retained ~30% of FTIR features	PLS feature reduction + SVM regression/classification	Test sets plus batch-wise external validation	[69]
40	Extra virgin olive oil	Detection of ternary adulteration with canola, hazelnut and safflower oils	ATR-FTIR	ATR correction; Savitzky–Golay smoothing/2nd derivative; spectral cutting; mean centering/Pareto scaling	Latent-variable modelling; comparison of scaling strategies; SIMCA class modelling	PCA; PLS-DA; SIMCA	Cross-validation + permutation testing + four external validation sets	[70]
41	Virgin coconut oil	Paraffin-oil adulteration detection and quantification	ATR-FTIR	Savitzky–Golay smoothing; mean centering + SNV; normal, 1st and 2nd derivatives; spectral-region optimization	PCA loading-plot selection of informative wavelengths; optimized spectral regions	PCA; LDA; PCR; PLS-R	Cross-validation plus genuinely external brand-based prediction set	[71]
42	Iranian rice	Variety authentication and adulteration quantification	FTIR	Autoscaling (mean centering / standard deviation scaling) of spectra	Embedded sparse variable selection via sLDA, lasso and elastic net; VIP-PLS comparator	sLDA; LDA; PCA-DA; PLS-DA; lasso; ridge; elastic net; VIP-PLS	70/30 external test set with comparative model evaluation	[72]
43	Cinnamon spice	Adulteration/replacement fraud detection	FT-IR	FT-IR spectral preprocessing used before PCA/classification; exact pretreatment final QC	No dedicated feature-selection algorithm central	PCA; PLS-DA; SIMCA	70/30 calibration-validation split with source samples kept independent between sets	[73]
44	Oregano	Adulteration detection and inter-laboratory transferability	Laboratory FT-IR comparator plus handheld NIR	Transferability/standardization preprocessing central to portable-device workflow	No dedicated feature selection central	RF for handheld NIR; accredited FT-IR chemometric reference method	Validation set plus international inter-laboratory transfer study	[74]
45	Black pepper	Geographical-origin authentication	FTIR and HS-SPME-GC-MS	FTIR spectral processing; GC-MS volatile profiling	PLS-DA variable interpretation; GC-MS terpenoid contributors	PCA for GC-MS; PLS-DA for FTIR	Cross-validation and external prediction for FTIR model	[75]

Table 2. Characteristics of GC-MS-based food-authentication studies included in the review (n = 29).

No	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML models	Validation Strategy	Ref.
1	Vinegar (rice, grape, apple, persimmon)	Type classification / authenticity	HS-SPME-GC-MS	StandardScaler; GC-MS deconvolution/ identification workflow	RFECV using Random Forest	XGBoost; KNN; RF; SVM; MLP	Stratified 80/20 hold-out; 5-fold GridSearchCV; RFECV-CV	[76]
2	Beef, pork and rabbit meat	Species discrimination / meat authentication	HS-SPME-GC-MS	Baseline correction; RI-based peak alignment; TIC normalization; missing-value imputation; autoscaling	Not clearly reported as a dedicated selection step	RF; XGBoost; ANN; t-SNE	Stratified 3-fold cross-validation	[77]
3	Wolfberry pulp (Lycium barbarum)	Geographical-origin authentication	HS-SPME-GC-MS plus GC-IMS/e-nose	GC-MS/volatile-data processing; scaling details not consistently explicit in retrieved text	VIP-based differential VOC selection	PCA; OPLS-DA; SVM; RF and related classifiers	70/30 train-test; cross-validation/permutation assessment for supervised models	[78]
4	Winter jujube	Geographical-origin authentication	GC-MS	GC-MS peak/compound matrix preparation; scaling details not prominent	Recursive Feature Elimination (RFE)	PLS-DA; SVM; RF/other ML classifiers	70/30 train-test; Leave-One-Out CV	[79]
5	Black tea	Geographical-origin authentication	HS-GC-MS	Full-spectrum/compound-profile data preparation	No dedicated feature-selection method clearly reported	kNN; RF	70/30 train-test; 5-fold GridSearchCV	[80]
6	Monofloral honey	Floral-origin authentication	HS-SPME-GC-MS	Normalized GC-MS peak-area ratios; scripted preprocessing	PLS-DA loadings; RF/SHAP importance	PLS-DA; RF; SHAP interpretation	Stratified 5-fold CV at sample level with replicates collapsed	[81]
7	Extra virgin olive oil	Cultivar/geographical-origin authentication	HS-SPME-GC-MS	Volatile-profile preprocessing/normalization	Variable-importance approach associated with tree-based modelling	RF and other classifiers	Cross-validation / held-out evaluation reported	[82]
8	Goose barnacle edible peduncle	Geographical traceability/authentication	GC-MS	GC-MS profile preparation; detailed normalization not fully specified in the extracted methods	No dedicated feature-selection method clearly identified in current extraction	Random Forest and multivariate analysis	Model validation reported; exact CV design not fully specified in the extracted methods	[83]
9	Dried ginger	Varietal and geographical-origin authentication	HS-GC-MS plus fast GC e-nose	Volatile-profile processing; GC-MS and e-nose handled separately	VIP >1 for key differential components	PCA/PLS-DA and several classifiers including RF	70/30 train-test; stratified 5-fold GridSearchCV	[84]

No	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML models	Validation Strategy	Ref.
10	Plant-based milk alternatives	Type/authenticity classification	HS-SPME-GC-MS	GC-MS preprocessing; normalization/scaling within workflow	RF feature importance	SVM; RF; kNN	Calibration + cross-validation + 32 additional biological validation samples	[85]
11	Coffee	Geographical-origin authentication	Targeted GC-MS/MS metabolomics	Derivatisation and targeted metabolite-profile processing; normalization/scaling reported in workflow	PLS-DA VIP + RF mean decrease Gini; intersection-feature strategy	PLS-DA; RF and other supervised ML models	Cross-validation plus 20% validation dataset	[86]
12	Lard, beef, chicken and mutton fats	Animal-fat discrimination / lard authentication	GC-MS n-alkane profiling	Normalization before chemometric modelling	VIP scores and RF variable importance; C24 proposed as marker	PCA; HCA; PLS-DA; RF	Normalized data split 90% training / 10% test; PLS-DA CV and 1000-permutation test	[87]
13	Candied Prunus mume	Geographical-origin differentiation	GC-MS volatile profiling plus untargeted metabolomics	VOC/metabolite profile processing; ROAV and network integration	RF-SHAP prioritization; WCNA modules	RF-SHAP; multivariate/network analyses	Model assessment reported; no clearly independent external cohort identified	[88]
14	Garlic	Geographical-origin authentication	GC-MS untargeted metabolomics	GC-MS metabolomics data processing followed by scaling/chemometric analysis	Discriminant metabolites/biomarker candidates from multivariate models	PLS-DA; OPLS-DA	Model validation/performance reported; exact external design limited	[89]
15	Salted goose	Breed/authenticity discrimination	HS-SPME-GC-MS plus GC-IMS, e-nose, e-tongue and sensory data	Platform-specific processing; GAN-based data augmentation and data fusion	Deep representation/data fusion rather than conventional univariate feature selection	GAN-integrated framework with RF, SVM, XGBoost and other base classifiers	80/20 train-test; cross-validation for hyperparameter selection	[90]
16	Coffee/chicory/soluble coffee	Coffee classification, origin/species discrimination and adulteration quantification	HS-SPME-GC-MS	GC-MS fingerprint/profile matrix processing; autoscaling for supervised models	Fingerprinting/profiling descriptors; no single tree-based selection method central	PCA; PLS-DA; PLS regression	Calibration/prediction splits; cross-validation	[20]
17	Safflower seed oil	Adulterant-concentration prediction	GC-MS fatty-acid measurements + hyperspectral imaging	Median filtering for hyperspectral data; GC-MS fatty-acid measurements integrated	Hyperspectral feature-band selection; GC-MS contributes selected fatty-acid variables	Ridge; PLSR; LightGBM stacking	5-fold cross-validation	[91]

No	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML models	Validation Strategy	Ref.
18	Olive oil	Vegetable-oil adulteration detection	HT-GC-MS (plus FIA-HESI-HRMS and HPLC-RID comparators)	Untargeted triacylglycerol fingerprint processing	Pattern-recognition variables from TAG fingerprints; no tree-based selector central	PLS-DA	Explicit external validation	[92]
19	Cocoa beans and cocoa liquors	Country-of-origin identification	HS-SPME-GC-MS	GC-MS fingerprint/profile processing; scaling/normalization used in chemometric workflow	Targeted profiling versus untargeted fingerprinting; discriminant variables interpreted	PCA; PLS-DA; OPLS-DA	Internal CV plus external test set for origin models	[93]
20	Virgin olive oil	Cultivar/varietal authentication	HS-SPME-GC-MS	Validated VOC quantitation/profile matrix; standard chemometric preparation	ANOVA selected discriminant VOCs from 72 quantified compounds	PLS-DA; LDA; hierarchical clustering	Internal modelling plus external validation	[94]
21	Virgin olive oil	Varietal authentication, especially Arbequina	HS-SPME-GC-MS	Chromatographic alignment/fingerprinting	Sesquiterpene fingerprint variables; model coefficients/markers interpreted	PLS-DA	Internal and external validation across broad agronomic/geographic conditions	[95]
22	Extra virgin/virgin olive oil	Geographical-origin authentication	HS-SPME-GC-MS	Aligned chromatographic fingerprint/profiling matrices	Targeted profiling versus untargeted fingerprinting	PLS-DA	Leave-10%-out cross-validation	[96]
23	Virgin olive oil	EU/non-EU and country geographical label authentication	HS-SPME-GC-MS	Sesquiterpene chromatographic fingerprint/alignment	Fingerprint variables used in PLS-DA	PLS-DA	External validation, three iterations	[97]
24	Hazelnut	Geographical and varietal authentication	GC-MS unsaponifiable fraction	Fingerprinting in selected-ion mode versus untargeted full-scan profiling	PLS-DA coefficients/high-relevance variables; discriminant compounds tentatively identified	PLS-DA	External validation	[98]
25	Extra virgin olive oil	PDO geographical-label compliance	HS-SPME-GC-MS	Sesquiterpene chromatographic fingerprint	Fingerprint/discriminant variables in DA	Discriminant analysis	External validation, three iterations	[99]
26	Virgin olive oil	Worldwide geographical-origin authentication	HS-SPME-GC-MS	Quantitative VOC profile preparation	One-way ANOVA selected 25 of 72 volatile compounds	LDA and alternative chemometric approaches	Internal validation on training plus independent external test	[100]

No	Food Matrix	Authentication Objective	Analytical Technique	Preprocessing	Feature / Variable Selection	Chemometric / ML models	Validation Strategy	Ref.
27	Commercial white wine	Grape-variety authentication/discrimination	Untargeted GC-MS	Peak deconvolution; baseline thresholding; S/N filtering; metabolite identification; retention-time/spectral alignment; scaling before multivariate modelling	Univariate screening and PLS-DA classifier-panel selection	ANOVA; PCA; ASCA; PLS-DA	Calibration modelling with independent test set	[101]
28	Oregano	Origin, variety and adulteration authentication	GC-MS plus MIR/NIR/HSI/PTR-TOF-MS	Technique-specific preprocessing; normalization	RF Gini-importance embedded selection	PCA; GP-LVM; nearest-neighbour; RF classifier/regressor	Repeated 70/30 train-test for classification; repeated leave-out regression	[102]
29	Rice	Geographical-origin and variety authentication	GC-MS plus NIR/MIR/HSI/PTR-TOF-MS	Peak detection/normalization; scaling to [0,1]	RF Gini importance; reduced feature sets (40 GC-MS; 20 NIR/MIR) before GP-LVM	GP-LVM + nearest-neighbour; RF used for feature ranking	Repeated 70/30 train-test procedure (20 repeats)	[103]

RQ1: Preprocessing Strategies in Food Authentication Based on FTIR and GC-MS

Preprocessing practices differed between FTIR- and GC-MS-based workflows. Among the 45 FTIR-based studies, the most frequently reported procedures were spectral derivatives (n = 13, 28.9%), SNV and Savitzky–Golay filtering or smoothing (each n = 11, 24.4%), mean-centering (n = 8, 17.8%), and MSC/EMSC (n = 7, 15.6%). Among the 29 GC-MS-based studies, scaling, autoscaling, or StandardScaler was most common (n = 11, 37.9%), followed by normalization (n = 10, 34.5%) and retention-time alignment (n = 5, 17.2%); peak deconvolution was reported in two studies (6.9%). The distribution of these preprocessing strategies is summarized in Table 3. Categories were not mutually exclusive because individual studies could combine multiple procedures.

Table 3. Distribution of preprocessing strategies reported in the included studies.

Preprocessing strategy	All studies, n (%)	FTIR-based, n (%)	GC-MS-based, n (%)
Baseline correction	6 (8.1%)	4 (8.9%)	2 (6.9%)
Savitzky–Golay / smoothing	11 (14.9%)	11 (24.4%)	0 (0.0%)
Derivative spectra	13 (17.6%)	13 (28.9%)	0 (0.0%)
SNV	11 (14.9%)	11 (24.4%)	0 (0.0%)
MSC/EMSC	7 (9.5%)	7 (15.6%)	0 (0.0%)
Mean-centering	8 (10.8%)	8 (17.8%)	0 (0.0%)
Scaling / autoscaling / StandardScaler	18 (24.3%)	7 (15.6%)	11 (37.9%)
Pareto scaling	2 (2.7%)	2 (4.4%)	0 (0.0%)
Normalization	13 (17.6%)	3 (6.7%)	10 (34.5%)
Spectral-region selection	4 (5.4%)	4 (8.9%)	0 (0.0%)
Retention-time alignment	5 (6.8%)	0 (0.0%)	5 (17.2%)
Peak deconvolution	2 (2.7%)	0 (0.0%)	2 (6.9%)
Transformation / log transformation	2 (2.7%)	2 (4.4%)	0 (0.0%)

Evidence on preprocessing effects was less consistent than usage frequencies. Goyal et al. [35] compared raw FTIR spectra with Savitzky–Golay smoothing, SNV, and EMSC to detect coriander powder adulteration with sawdust, whereas Tarandakzad et al. [36] evaluated a broader range of procedures, including scaling, scatter correction, smoothing, derivatives, wavelet processing, and median filtering. However, the included studies did not provide sufficiently consistent, comparable performance metrics to establish a quantitative ranking of preprocessing methods. Frequency of use therefore should not be interpreted as evidence of superior predictive performance. Identifying an optimal method requires controlled comparisons using the same dataset, prediction task, model, and validation design. Accordingly, preprocessing should be treated as a model-dependent choice and evaluated using appropriate validation procedures rather than assumed to be universally optimal.

RQ2: Strategy of Feature Selection and Variables

The reported approaches to feature and variable selection were heterogeneous. Following a full-text audit of the ten studies initially grouped under VIP/loading-based approaches, four studies explicitly used selected variables or wavenumbers in a subsequent predictive model, whereas six primarily used VIP, loading, or related importance measures to identify discriminant variables or biomarkers without sufficient evidence that a reduced model was retrained and independently evaluated. Other reported approaches included RFE/RFECV (n = 5), univariate statistical selection (n = 4), and spectral interval/region selection (n = 3). Twenty studies (27.0%) explicitly reported no dedicated feature-selection procedure, while 24 (32.4%) were classified as other/unclear because their approaches were heterogeneous or insufficiently described. The distribution of these approaches and their classification is presented in Table 4.

Table 4. Primary feature and variable-selection strategies, distinguished from post-model variable interpretation and biomarker discovery.

Primary feature-/variable-selection strategy	Studies	% of 74
VIP/loading used in a subsequent predictive model	4	5.4%
VIP/loading/importance used for interpretation or biomarker discovery	6	8.1%
RFE/RFECV	5	6.8%
Univariate statistical selection	4	5.4%
Spectral interval/region selection	3	4.1%
Tree/RF importance (without verified reduced-model evaluation)	2	2.7%
Latent-variable reduction	2	2.7%
Embedded/deep-learned features	2	2.7%
CARS	1	1.4%
Sparse/penalized selection	1	1.4%
No dedicated feature selection	20	27.0%
Other/unclear	24	32.4%

In FTIR studies, selection commonly targeted informative wavenumbers or spectral regions, whereas GC-MS studies more often targeted discriminatory metabolites, volatile compounds, fatty acids, or related chemical markers. For this review, VIP scores, loadings, SHAP values, and Random Forest importance were not treated as proof of predictive feature selection by themselves. They were classified as predictive selection only when the selected variables were explicitly carried forward into a subsequent model; otherwise, they were treated as variable interpretation or biomarker discovery. Even when a reduced model was reported, the strength of evidence depended on the validation design and whether selection was performed without information leakage.

RQ3: Validation Strategies and Model Generalization Capabilities

The validation designs varied considerably across the 74 included studies. External or independent validation was identified in 21 studies (28.4%), internal hold-out testing in 27 studies (36.5%), cross-validation only in 13 studies (17.6%), exploratory analyses without predictive validation in one study (1.4%), and vague or other validation designs in 12 studies (16.2%). Although 21 studies were initially classified as using external or independent validation, these designs should not be assumed to provide equivalent evidence of model generalizability. A separate test set can evaluate model performance on unseen samples, but stronger evidence of transferability requires clear information about the independence of the samples and the conditions under which they were acquired, such as different sources, batches, locations, acquisition periods, laboratories, or instruments. These details were not consistently established in the extracted data; therefore, the proportion of studies classified as external or independent validation should be interpreted cautiously.

A separate audit of information-leakage control revealed an important methodological limitation across the included studies. Only three of the 74 studies (4.1%) explicitly documented procedures intended to prevent information leakage. Six studies (8.1%) raised potential leakage concerns, whereas the remaining 65 studies (87.8%) provided unclear or insufficient information to determine whether leakage had been adequately controlled. These categories reflect the evidence available in the reviewed reports and should not be interpreted as proof that leakage occurred in every study with insufficient reporting.

The three studies with explicitly documented leakage-control procedures were Park et al. [76], Pirhadi et al. [41], and Ennahli et al. [43]. The reported safeguards included restricting resampling and feature-selection procedures to the training workflow, fitting preprocessing

transformations within training folds, and partitioning data at the biological-sample level to prevent technical replicates from the same sample from appearing in both training and test sets. These practices illustrate that reliable validation requires more than a nominal train–test split: data-dependent preprocessing, feature selection, resampling, and model optimization must be conducted without using information from held-out observations.

Importantly, external validation and leakage control represent distinct methodological dimensions. An independent test set may provide evidence of performance on unseen samples, but it does not by itself establish that preprocessing, feature selection, and hyperparameter tuning were conducted without access to test-set information. Conversely, leakage-aware internal validation does not necessarily demonstrate transferability across producers, batches, locations, seasons, laboratories, instruments, or acquisition periods

The handling of technical replication represents an important methodological consideration in model validation. Ennahli et al. [43] for example, partitioned 65 biological honey samples before model development, ensuring that the five technical replicates associated with each biological sample were not distributed across both the training and test sets. This design reduces the risk of overly optimistic performance estimates caused by related measurements appearing in both sets. Li et al. [49] also reported sample-level handling and an independent testing structure in a multiplatform tea-authentication study. Such designs provide stronger evidence of generalization to unseen biological samples than random partitioning at the spectrum level, where measurements from the same sample may occur in both training and test sets. Nevertheless, sample-level independence alone does not establish generalizability across laboratories, instruments, or acquisition conditions.

Tree-based Modelling in FTIR and GC-MS Food Authentication

Tree-based algorithms were found in 31 of 74 studies (41.9%), but the available evidence is dominated by the ensemble method. Random Forest was used in 27 studies (36.5%), whereas single Decision Tree was found in only three studies (4.1%), and the boosting/bagging approach in seven studies (9.5%). These categories are not mutually exclusive. Random Forest is used both as a predictive model and as a mechanism for determining variable importance.

Therefore, the systematic evidence supports the use of the term tree-based modeling rather than implying that the standalone Decision Tree is the dominant approach. Performance on external validation may also be lower than internal performance, even when feature reduction has been performed. This pattern is evident in the meat authentication and quality study by Fengou et al. [69], in which the FTIR-based model achieved internal test-set accuracies of 99%, 96%, 100%, and 97% across four batches (b1–b4), compared with external validation accuracies of 94%, 88%, 92%, and 92%, respectively. These results represent absolute decreases of 5–8 percentage points across the external batches. Although external accuracy remained relatively high, the consistent reduction highlights that strong internal performance does not necessarily translate into equivalent performance on independent samples. Accordingly, model performance should be interpreted in relation to the independence of the evaluation data and the model's ability to generalize beyond the data used during model development, rather than on internal accuracy alone

Integrated Findings Across RQ1-RQ3

Based on the three review questions, preprocessing, feature selection, model development, and validation emerge as interdependent components of a single analytical pipeline. Although high classification accuracy and strong regression statistics were widely reported across the included studies, the reliability of these performance estimates depends on several methodological factors, including sample independence, replication handling, the timing of preprocessing and feature selection, and whether the final model was evaluated using data independent of model

optimization. These findings highlight the importance of considering the entire analytical workflow rather than interpreting predictive performance metrics in isolation. Table 5 presents an integrated methodological synthesis of preprocessing, feature-selection, model-development, and validation practices, highlighting their implications for the reliability and generalizability of food authentication models.

Table 5. Integrated methodological synthesis of preprocessing, feature-selection, and validation practices.

Methodological Component	Main Pattern Identified	Main Methodological Concern	Implication for Model Development
FTIR preprocessing	Derivatives, SNV, SG, MSC/EMSC and mean-centering were commonly reported.	No universally optimal preprocessing; sequence and parameter choice varied.	Compare preprocessing within the training/resampling framework.
GC-MS preprocessing	Scaling, normalization and retention-time alignment were prominent.	Peak-processing workflows were heterogeneous and incompletely reported in some studies.	Report peak processing, alignment, normalization and their sequence explicitly.
Feature selection	VIP/loading, RFE/RFECV, RF importance, interval selection and other methods were used.	Selection timing relative to resampling was frequently unclear.	Perform data-dependent feature selection within training/CV folds.
Internal validation	Hold-out tests and cross-validation were widely used.	Internal performance may overestimate transfer to new samples.	Separate optimization from final untouched evaluation.
External validation	External/independent validation was identified in 21/74 studies.	Evidence of transferability remained limited across the literature.	Use independent batches, origins, producers, brands, seasons or acquisition periods when feasible.
Data-leakage control	Only 6/74 studies explicitly documented strong leakage-control procedures.	Optimistic performance cannot always be ruled out when workflow timing is unclear.	Fit preprocessing, feature selection and tuning using training data only.
Tree-based modelling	Tree-based methods appeared in 31/74 studies and were dominated by Random Forest.	Standalone Decision Tree evidence was sparse, and high internal accuracy does not establish generalization.	Evaluate tree-based models with leakage-aware validation and independent testing.

Implications for Future FTIR and GC-MS-based Food Authentication Research

Future studies should report the complete sequence of preprocessing, feature selection, dimensionality reduction, model fitting, hyperparameter optimization, and validation. Sample-size reporting should also clearly distinguish independent biological samples from technical spectra or repeated chromatographic injections. External validation should approximate real-world application scenarios as closely as possible, for example by using different producers, locations, harvest seasons, processing batches, brands, instruments, or acquisition periods. Feature selection should be evaluated not only by improvements in accuracy but also by stability and chemical interpretability. Finally, comparisons across algorithms should prioritize generalizability rather than maximum internal performance.

Strengths and Limitations Review

One of the strengths of this review is the integration of preprocessing, feature selection, validation, data leakage reporting, and tree-based modeling into one systematic evidence

framework. A total of 74 studies remained in review despite not using feature selection, Decision Tree, or external validation. This approach allows the absence of such methodological characteristics to be synthesized, rather than excluded from the review. The review also makes a clear distinction between internal hold-out testing and external validation/independence, as well as a distinction between vague data leakage reporting and confirmed data leakage incidents.

However, several limitations should be considered when interpreting the findings of this review. First, the systematic search was restricted to ScienceDirect, meaning that relevant studies indexed in other bibliographic databases may have been missed. In addition, the inclusion of model- and validation-specific terms, such as “Random Forest” and “validation,” may have favored the retrieval of studies explicitly reporting these methods in their searchable metadata. Consequently, the observed frequencies of modeling and validation strategies describe the included corpus and should not be interpreted as estimates of their prevalence across the broader food-authentication literature. Second, 30 of the 115 reports sought for retrieval could not be obtained for full-text assessment. Because their eligibility and methodological characteristics could not be verified, it remains unclear whether their exclusion affected the observed distribution of analytical methods and validation strategies.

This may have introduced retrieval bias and limits the representativeness of the included corpus. Accordingly, the findings should be interpreted as a synthesis of the 74 included studies rather than a comprehensive representation of all potentially eligible studies.. Third, the included studies were heterogeneous in terms of food matrices, sample sizes, analytical conditions, algorithms, and performance metrics, limiting the feasibility of meaningful quantitative meta-analysis. Fourth, incomplete reporting in some primary studies limited the assessment of preprocessing sequences, the timing of feature selection, and information-leakage prevention. Fifth, the methodological-quality appraisal was developed specifically for this review and has not been validated as a formal risk-of-bias instrument. Finally, AI tools supported data filtering and extraction within an author-verified workflow; however, this assistance does not constitute independent verification by a second human reviewer. Future reviews should consider searching multiple bibliographic databases and using complementary search strategies to improve coverage and assess the robustness of the findings.

Conclusions

This systematic critical review of 74 studies demonstrates that the reliability of FTIR- and GC-MS-based food authentication models depends on the integration of preprocessing, feature selection, model development, and validation. No preprocessing or feature-selection strategy was universally optimal, while Random Forest was more frequently represented than standalone Decision Tree models within the reviewed corpus. External or independent validation was reported in 21 of the 74 studies, while incomplete methodological reporting often limited verification of information-leakage control. Future studies should adopt leakage-aware workflows in which data-dependent preprocessing, feature selection, and model optimization are performed exclusively within training data, including each training fold during cross-validation. Researchers should also report the unit of independence used for data partitioning, distinguish independent biological samples from the numbers of spectra or chromatographic injections, and clearly describe how external validation samples are independent of model development. These practices are essential to improve reproducibility and provide more credible assessments of model generalizability.

Funding: This research was funded by the Kementerian Pendidikan Tinggi, Sains, dan Teknologi through the Program Penelitian Disertasi Doktor, master contract number: 271/C3/DT.05.00/PLBARU/2026, and subcontract number: 170/UN11.L1/PG.01.03/D PPM/2026.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

[1] Brooks C, Parr L, Smith JM, Buchanan D, Snioch D, Hebishy E. A Review of Food Fraud and Food Authenticity across the Food Supply Chain, with an Examination of the Impact of the COVID-19 Pandemic and Brexit on Food Industry. *Food Control* 2021;130:108171. https://doi.org/10.1016/j.foodcont.2021.108171.

[2] Esteki M, Regueiro J, Simal-Gándara J. Tackling Fraudsters with Global Strategies to Expose Fraud in the Food Chain. *Comprehensive Reviews in Food Science and Food Safety* 2019;18:425–40. https://doi.org/10.1111/1541-4337.12419.

[3] Kwasi Bannor R, Arthur KK, Oppong D, Oppong-Kyeremeh H. A Comprehensive Systematic Review and Bibliometric Analysis of Food Fraud from a Global Perspective. *Journal of Agriculture and Food Research* 2023;14:100686. https://doi.org/10.1016/j.jafr.2023.100686.

[4] Visciano P, Schirone M. Food Frauds: Global Incidents and Misleading Situations. *Trends in Food Science & Technology* 2021;114:424–42. https://doi.org/10.1016/j.tifs.2021.06.010.

[5] Vicente J, Kist V, Mega CP, Rossato OB, de Oliveira Veloso G, de Carvalho MG. Spectroscopy-Driven Authentication of Agro-Industrial Products: A Comprehensive Review of NMR, FT-IR, NIR and Chemometric Workflows. *Food Chemistry* 2026;513:148982. https://doi.org/10.1016/j.foodchem.2026.148982.

[6] Quintanilla-Casas B, Torres-Cobos B, Bro R, Guardiola F, Vichi S, Tres A. The Volatile Metabolome — Gas Chromatography–Mass Spectrometry Approaches in the Context of Food Fraud. *Current Opinion in Food Science* 2025;61:101235. https://doi.org/1016/j.cofs.2024.101235.

[7] Rohman A, Ghazali MAB, Windarsih A, Irnawati, Riyanto S, Yusof FM, et al. Comprehensive Review on Application of FTIR Spectroscopy Coupled with Chemometrics for Authentication Analysis of Fats and Oils in the Food Products. *Molecules* 2020;25. https://doi.org/10.3390/molecules25225485.

[8] Biancolillo A, Marini F, Ruckebusch C, Vitale R. Chemometric Strategies for Spectroscopy-Based Food Authentication. *Applied Sciences* 2020;10. https://doi.org/10.3390/app10186544.

[9] González-Domínguez R, Sayago A, Fernández-Recamales Á. An Overview on the Application of Chemometrics Tools in Food Authenticity and Traceability. *Foods* 2022;11. https://doi.org/10.3390/foods11233940.

[10] Granato D, Putnik P, Kovačević DB, Santos JS, Calado V, Rocha RS, et al. Trends in Chemometrics: Food Authentication, Microbiology, and Effects of Processing. *Comprehensive Reviews in Food Science and Food Safety* 2018;17:663–77. https://doi.org/10.1111/1541-4337.12341.

[11] Castañeda FN, Pellegrino Vidal RB, Aspromonte J. Untargeted Chromatographic Methods Coupled with Chemometric Strategies for the Analysis of Food and Related Samples. *TrAC Trends in Analytical Chemistry* 2024;173:117650. https://doi.org/10.1016/j.trac.2024.117650.

[12] Zhang Y, Jin L, Guo F, Ni X, Zhao Y, Cheng Y, et al. Matrix Factorization-Based Dimensionality Reduction Algorithms-A Comparative Study on Spectroscopic Profiling Data. *Analytical Chemistry* 2022;94:13385–95. https://doi.org/10.1021/acs.analchem.2c01922.

[13] Christmann J, Rohn S, Weller P. Finding Features - Variable Extraction Strategies for Dimensionality Reduction and Marker Compounds Identification in GC-IMS Data. *Food Research International* 2022;161:111779. https://doi.org/10.1016/j.foodres.2022.111779.

[14] Chahal S, Tian L, Bilamjian S, Balogh F, De Leoz L, Anumol T, et al. Robust Multiclass Feature Selection for the Authentication of Honey Botanical Origin via Nontargeted LC-MS Analysis. *Analytical Chemistry* 2025;97:12521–30. https://doi.org/10.1021/acs.analchem.4c06723.

[15] Yu H-C, Chen Y-K. Improved FTIR-Based Classification for Food Authentication Using a Topological Ensemble Framework. *Current Research in Food Science* 2026;12:101441. https://doi.org/10.1016/j.crfs.2026.101441.

[16] Yan C. A Review on Spectral Data Preprocessing Techniques for Machine Learning and Quantitative Analysis. *IScience* 2025;28. https://doi.org/10.1016/j.isci.2025.112759.

[17] Pérez-Cova M, Tauler R, Jaumot J. Chemometrics in Comprehensive Two-Dimensional Liquid Chromatography: A Study of the Data Structure and Its Multilinear Behavior. *Chemometrics and Intelligent Laboratory Systems* 2020;201:104009. https://doi.org/10.1016/j.chemolab.2020.104009.

[18] Deryck A, Fernández Pierna JA, Baeten V, Lesnoff M. A Step-by-Step Tutorial to Local Partial Least Squares Analyses for near-Infrared Spectroscopic Data Analysis. *Analytica Chimica Acta* 2026;1401:345242. https://doi.org/10.1016/j.aca.2026.345242.

[19] Folli GS, de Paulo EH, Santos FD, Nascimento MHC, da Cunha PHP, Romão W, et al. Correlation Analysis of Modern Analytical Data – a Chemometric Dissection of Spectral and Chromatographic Variables. *Analytical Methods* 2023;15:4119–33. https://doi.org/10.1039/d3ay00711a.

[20] Núñez N, Moret E, Lucci P, Moret S, Saurina J, Núñez O. SPME-GC–MS and Chemometrics for Coffee Characterization, Classification and Authentication. *Microchemical Journal* 2025;213:113771. https://doi.org/10.1016/j.microc.2025.113771.

[21] Mészáros M, Sedlák J, Bílek T, Vávra A. Evaluating LDA and PLS-DA Algorithms for Food Authentication: A Chemometric Perspective. *Algorithms* 2025;18. https://doi.org/10.3390/a18120733.

[22] Rodionova OY, Oliveri P, Malegori C, Pomerantsev AL. Chemometrics as an Efficient Tool for Food Authentication: Golden Pillars for Building Reliable Models. *Trends in Food Science & Technology* 2024;147:104429. https://doi.org/10.1016/j.tifs.2024.104429.

[23] Bi Y, Yuan K, Xiao W, Wu J, Shi C, Xia J, et al. A Local Pre-Processing Method for near-Infrared Spectra, Combined with Spectral Segmentation and Standard Normal Variate Transformation. *Analytica Chimica Acta* 2016;909:30–40. https://doi.org/10.1016/j.aca.2016.01.010.

[24] Jiao Y, Li Z, Chen X, Fei S. Preprocessing Methods for Near-Infrared Spectrum Calibration. *Journal of Chemometrics* 2020;34:e3306. https://doi.org/10.1002/cem.3306.

[25] Silalahi DD, Midi H, Arasan J, Mustafa MS, Caliman J-P. Robust Generalized Multiplicative Scatter Correction Algorithm on Pretreatment of near Infrared Spectral Data. *Vibrational Spectroscopy* 2018;97:55–65. https://doi.org/10.1016/j.vibspec.2018.05.002.

[26] Munawar AA, Hizir, Erika C, Pawelzik E. Fast and Simultaneous Prediction of Inner Quality Parameters on Intact Mangos by near Infrared Spectroscopy: Impact of Spectra Pre-Processing on Prediction Accuracy. *Future Foods* 2024;10:100463. https://doi.org/10.1016/j.fufo.2024.100463.

[27] Mishra P, Lohumi S. Improved Prediction of Protein Content in Wheat Kernels with a Fusion of Scatter Correction Methods in NIR Data Modelling. *Biosystems Engineering* 2021;203:93–7. https://doi.org/biosystemseng.2021.01.003.

[28] De Araújo Gomes A, Azcarate S, Diniz P, De Sousa Fernandes DD, Véras G. Variable Selection in the Chemometric Treatment of Food Data: A Tutorial Review. *Food Chemistry* 2021;370:131072. https://doi.org/10.1016/j.foodchem.2021.131072.

[29] Smolinska A, Pellacani S, Skawinski M, Durante C. On the Road to Automation: A Comparative Review on Chemometric Strategies for GC-MS Data Analysis. *TrAC Trends in Analytical Chemistry* 2025;190:118286. https://doi.org/10.1016/j.trac.2025.118286.

[30] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 2021:n71. https://doi.org/10.1136/bmj.n71.

[31] Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: An Extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews* 2021;10:39.

[32] Wu X, Wang J, Zhao X, Zhao X, Wang S, Zhang Y. Optimization of Convenient Infrared Spectroscopy Detection Methods and Infrared-Raman Fusion Spectroscopic Detection for Milk Powder Adulteration. *Journal of Food Composition and Analysis* 2025;148:108195. https://doi.org/10.1016/j.jfca.2025.108195.

[33] Lima JS, Ribeiro DCSZ, Neto HA, Campos SVA, Leite MO, Fortini ME de R, et al. A Machine Learning Proposal Method to Detect Milk Tainted with Cheese Whey. *Journal of Dairy Science* 2022;105:9496–508. https://doi.org/10.3168/jds.2021-21380.

[34] Vignesh T, Selvakumar D, Mallikarjun PY, Geetha NK, Madhankumar S, Jatkar M. FTIR Spectroscopy and Machine Learning for Detection of Sudan I Dye as a Hazardous Contaminant in Turmeric Powder. *Journal of Environmental Chemical Engineering* 2026;14:123709. https://doi.org/10.1016/j.jece.2026.123709.

[35] Goyal R, Singha P, Singh SK. Machine Learning-Assisted Fourier Transform Infrared Spectroscopy to Predict Adulteration in Coriander Powder. *Food Chemistry* 2025;477:143502. https://doi.org/10.1016/j.foodchem.2025.143502.

[36] Tarandakzad Y, Khodabakhshian R, Khojastehpour M, Mahdavianmehr H. FTIR Spectroscopy Identification of Fraud in Coffee Powder: A Study on Preprocessing Methods. *Food Chemistry* 2026;518:148777. https://doi.org/10.1016/j.foodchem.2026.148777.

[37] Aliaño-González MJ, Vázquez-Espinosa M, V. González-de-Peredo A, Chinchilla N, Fernández Barbero G, Palma M, et al. Machine Learning-Assisted Spectroscopic Methods for Detecting Adulteration in Barrantes

Wine from Folla Redonda Grapes. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 2026;344:126723. https://doi.org/10.1016/j.saa.2025.126723.

[38] Wang Y-T, Ren H-B, Liang W-Y, Jin X, Yuan Q, Liu Z-R, et al. A Novel Approach to Temperature-Dependent Thermal Processing Authentication for Milk by Infrared Spectroscopy Coupled with Machine Learning. *Journal of Food Engineering* 2021;311:110740. https://doi.org/10.1016/j.jfoodeng.2021.110740.

[39] Paulo EH d., Rech AM, Weiler FH, Nascimento MHC, Filgueiras PR, Ferrão MF. Evaluation of Adulteration in Soy-Based Beverages by Water Addition Using Chemometrics Applied to ATR-FTIR Spectroscopy. *Food Control* 2024;166:110746. https://doi.org/10.1016/j.foodcont.2024.110746.

[40] Dimakopoulou-Papazoglou D, Serrano S, Rodríguez I, Ploskas N, Koutsoumanis K, Katsanidis E. FTIR Spectroscopy Combined with Machine Learning for the Classification of Mediterranean Honey Based on Origin. *Journal of Food Composition and Analysis* 2025;144:107778. https://doi.org/10.1016/j.jfca.2025.107778.

[41] Pirhadi M, Shariatifar N, Pirhadi S, Jahed Khaniki G, Amanlou M, Sadighara P, et al. Classification of Meat Species in Sausage Products Using FTIR Spectroscopy and Multivariate Analysis for Food Authentication. *Results in Chemistry* 2025;18:102871. https://doi.org/10.1016/j.rechem.2025.102871.

[42] Hssaini L, Ennahli S. Preprocessing-Dependent Machine Learning Models Enhance Mid-FTIR Detection of Honey Adulteration. *Journal of Food Composition and Analysis* 2025;148:108666. https://doi.org/10.1016/j.jfca.2025.108666.

[43] Ennahli S, Ajal EA, Bajoub A, Hssaini L. Rapid Prediction of Honey Adulteration Using Machine Learning-Assisted FTIR Spectroscopy and Chemometrics. *Food Control* 2026;187:112146. https://doi.org/10.1016/j.foodcont.2026.112146.

[44] de Santana FB, Borges Neto W, Poppi RJ. Random Forest as One-Class Classifier and Infrared Spectroscopy for Food Adulteration Detection. *Food Chemistry* 2019;293:323–32. https://doi.org/10.1016/j.foodchem.2019.04.073.

[45] Rodrigues de Aguiar T, Dorocz EL, Dalagrana do Santos L, Coelho Tanamati AA, Gozzo AM, Bona E. Mid-Infrared Spectroscopy and Chemometrics in the Detection of Adulteration in Chia Oil (Salvia Hispanica L) and α-Linolenic Acid Content Prediction. *Food Control* 2024;165:110687. https://doi.org/10.1016/j.foodcont.2024.110687.

[46] Colak S, Uzunsoy I, Narin A, Duran U. Adulteration Detection of Cow Milk in Buffalo Milk Using Fourier-Transform Infrared Spectroscopy and Artificial Intelligence-Based Techniques. *Journal of Food Composition and Analysis* 2025;140:107203. https://doi.org/10.1016/j.jfca.2025.107203.

[47] Bergamaschi M, Cipolat-Gotet C, Cecchinato A, Schiavon S, Bittante G. Chemometric Authentication of Farming Systems of Origin of Food (Milk and Ripened Cheese) Using Infrared Spectra, Fatty Acid Profiles, Flavor Fingerprints, and Sensory Descriptions. *Food Chemistry* 2020;305:125480. https://doi.org/10.1016/j.foodchem.2019.125480.

[48] Jayasekher A, Panchariya PC, Maurelli F, Prajapati D, Palit AK. Authentication of Mentha Arvensis Essential Oil Using Attenuated Total Reflectance-Fourier Transform Infrared Spectrophotometry Coupled with Chemometrics. *Journal of Food Composition and Analysis* 2024;135:106576. https://doi.org/10.1016/j.jfca.2024.106576.

[49] Li Y, Logan N, Petchkongkaew A, Hong Y, Liu X, Birse N, et al. Transforming Food Authenticity Testing by the Exploitation of a Machine Learning – Data Fusion Approach: A Tea Case Study. *Food Chemistry* 2026;508:148302. https://doi.org/10.1016/j.foodchem.2026.148302.

[50] He W, Hao G, Wang S, Meng R, Wang D, Wang Y, et al. Machine Learning-Enhanced Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) Spectroscopy for Discrimination of Green Arabica and Robusta Coffee Beans. *Journal of Agriculture and Food Research* 2026;26:102689. https://doi.org/10.1016/j.jafr.2026.102689.

[51] Teklemariam TA, Moisey J, Gotera J. Attenuated Total Reflectance-Fourier Transform Infrared Spectroscopy Coupled with Chemometrics for the Rapid Detection of Coconut Water Adulteration. *Food Chemistry* 2021;355:129616. https://doi.org/10.1016/j.foodchem.2021.129616.

[52] En-Najy A, Kharbach M, Vander Heyden Y, Bouklouze A. Adulteration Assessment of Four High-Priced Moroccan Honeys with Artificial Honey Using ATR-FTIR Fingerprints in Tandem with Chemometric Tools. *Journal of Food Composition and Analysis* 2026;157:109386. https://doi.org/10.1016/j.jfca.2026.109386.

[53] Grabato JR, Pilario KE, Micor JRL, Mojica E-RE. Geographical and Entomological Differentiation of Philippine Honey by Multivariate Analysis of FTIR Spectra. *Journal of Food Composition and Analysis* 2022;114:104853. https://doi.org/10.1016/j.jfca.2022.104853.

[54] Khodabakhshian R, Seyedalibeyk Lavasani H, Weller P. A Methodological Approach to Preprocessing FTIR Spectra of Adulterated Sesame Oil. *Food Chemistry* 2023;419:136055. https://doi.org/10.1016/j.foodchem.2023.136055.

[55] Brar DS, Singh B, Nanda V. Transformative Detection of Sudan IV Adulteration in Prominent Indian Red Chilli Varieties: Chemometric Approach Coupling ATR-FTIR Spectroscopy and DD-SIMCA One Class Model. *Journal of Food Composition and Analysis* 2025;147:107995. https://doi.org/10.1016/j.jfca.2025.107995.

[56] Khanmohammadi Khorrami M, Ahmadvand Shahverdi M, Asadian M, Shirinnejad M, Mohammadi M, Shirian AZ, et al. Combining ACE, PLS-R, and SVM-R for Rapid Detection of Adulteration in Saffron Samples by Diffuse Reflectance Infrared Fourier Transform Spectroscopy. *Food Control* 2025;168:110853. https://doi.org/10.1016/j.foodcont.2024.110853.

[57] Jingying C, Yunqing Z, Loh JY, Low CY, Wujun Z, Yingzhen H, et al. Rapid and Accurate Classification of Dioscorea Species Using Attenuated Total Reflection-Fourier Transform Infrared (ATR-FTIR) Spectroscopy Integrated with Chemometric Analysis and Convolutional Neural Network (CNN). *Applied Food Research* 2026;6:102213. https://doi.org/10.1016/j.afres.2026.102213.

[58] Rozali NL, Azizan KA, Singh R, Syed Jaafar SN, Othman A, Weckwerth W, et al. Fourier Transform Infrared (FTIR) Spectroscopy Approach Combined with Discriminant Analysis and Prediction Model for Crude Palm Oil Authentication of Different Geographical and Temporal Origins. *Food Control* 2023;146:109509. https://doi.org/10.1016/j.foodcont.2022.109509.

[59] Zhang X, Wu X, Du Z, Yang D, Wang S, Wang J, et al. Discrimination of Edible Vegetable Oils from Different Processing Techniques and Rapeseed Oil Adulteration Using Fused Raman-FTIR Spectroscopy and 1D CNN. *Journal of Food Composition and Analysis* 2026;149:108785. https://doi.org/10.1016/j.jfca.2025.108785.

[60] Boateng AA, Sumaila S, Lartey M, Oppong MB, Opuni KFM, Adutwum LA. Evaluation of Chemometric Classification and Regression Models for the Detection of Syrup Adulteration in Honey. *LWT* 2022;163:113498. https://doi.org/10.1016/j.lwt.2022.113498.

[61] Rajagukguk YV, Tomaszewska-Gras J. Advanced Authenticity Screening of Commercial Berry Seed Oils Using Full FTIR Spectra and DSC Curves Coupled with Chemometrics. *LWT* 2025;222:117680. https://doi.org/10.1016/j.lwt.2025.117680.

[62] Cruz-Tirado JP, Haughey SA, Elliott CT, Barbin DF. Portable NIR Spectrometer and FTIR and Soft Modelling to Authenticate Edible Cricket, Mealworm and Buffalo Worm Flour. *Food Control* 2026;179:111567. https://doi.org/10.1016/j.foodcont.2025.111567.

[63] Halagarda M, Zaczyk M, Popek S, Pedan V, Kurczab R, Rohn S. Honey Differentiation with FTIR-ATR Spectroscopy – Comparison with Physicochemical Parameters of a Polish Honey Sample Set. *Journal of Food Composition and Analysis* 2024;130:106195. https://doi.org/10.1016/j.jfca.2024.106195.

[64] Julmohammad N, Tan E. ATR-FTIR with Chemometrics Tools for Classification and Prediction of Hydrogen Peroxide in Liquid Milk. *Journal of Food Composition and Analysis* 2025;140:107223. https://doi.org/10.1016/j.jfca.2025.107223.

[65] Amit, Kumari S, Jamwal R, Suman P, Singh DK. Expeditious and Accurate Detection of Palm Oil Adulteration in Virgin Coconut Oil by Utilizing ATR-FTIR Spectroscopy along with Chemometrics and Regression Models. *Food Chemistry Advances* 2023;3:100377. https://doi.org/10.1016/j.focha.2023.100377.

[66] Sitorus A, Bulan R. Development of a Screening Method for Adulteration Detection in Coconut Milk via Mid-Infrared Spectroscopy: A Study of Linear and Nonlinear Regression Method. *Journal of Agriculture and Food Research* 2022;10:100438. https://doi.org/10.1016/j.jafr.2022.100438.

[67] Ye Q, Meng X. Highly Efficient Authentication of Edible Oils by FTIR Spectroscopy Coupled with Chemometrics. *Food Chemistry* 2022;385:132661. https://doi.org/10.1016/j.foodchem.2022.132661.

[68] Rovira G, Miaw CSW, Martins MLC, Sena MM, de Souza SVC, Ruisánchez I, et al. In-Depth Chemometric Strategy to Detect up to Four Adulterants in Cashew Nuts by IR Spectroscopic Techniques. *Microchemical Journal* 2022;181:107816. https://doi.org/10.1016/j.microc.2022.107816.

[69] Fengou L-C, Lytou AE, Tsekos G, Tsakanikas P, Nychas G-JE. Features in Visible and Fourier Transform Infrared Spectra Confronting Aspects of Meat Quality and Fraud. *Food Chemistry* 2024;440:138184. https://doi.org/10.1016/j.foodchem.2023.138184.

[70] Ordoudi SA, Özdikicierler O, Tsimidou MZ. Detection of Ternary Mixtures of Virgin Olive Oil with Canola, Hazelnut or Safflower Oils via Non-Targeted ATR-FTIR Fingerprinting and Chemometrics. *Food Control* 2022;142:109240. https://doi.org/10.1016/j.foodcont.2022.109240.

[71] Amit, Jamwal R, Kumari S, Dhaulaniya AS, Balan B, Kelly S, et al. Utilizing ATR-FTIR Spectroscopy Combined with Multivariate Chemometric Modelling for the Swift Detection of Mustard Oil Adulteration in Virgin Coconut Oil. *Vibrational Spectroscopy* 2020;109:103066. https://doi.org/10.1016/j.vibspec.2020.103066.

[72] Rahmani N, Mani-Varnosfaderani A. Quality Control, Classification, and Authentication of Iranian Rice Varieties Using FT-IR Spectroscopy and Sparse Chemometric Methods. *Journal of Food Composition and Analysis* 2022;112:104650. https://doi.org/10.1016/j.jfca.2022.104650.

[73] Lixourgioti P, Goggin KA, Zhao X, Murphy DJ, van Ruth S, Koidis A. Authentication of Cinnamon Spice Samples Using FT-IR Spectroscopy and Chemometric Classification. *LWT* 2022;154:112760. https://doi.org/10.1016/j.lwt.2021.112760.

[74] McGrath TF, Haughey SA, Islam M, Elliott CT, Kelly SD, Suman M, et al. The Potential of Handheld near Infrared Spectroscopy to Detect Food Adulteration: Results of a Global, Multi-Instrument Inter-Laboratory Study. *Food Chemistry* 2021;353:128718. https://doi.org/10.1016/j.foodchem.2020.128718.

[75] Biancolillo A, Rossi M, Reale S, Scappaticci C, Foschi M, D'Archivio AA. Multi-Analytical Approach for the Characterization and Geographical Authentication of Black Pepper: Volatile Analysis, Antioxidant Activity, and Spectroscopic Profiling. *Microchemical Journal* 2026;221:116895. https://doi.org/10.1016/j.microc.2026.116895.

[76] Park S, Kim K, Sung J, Son H, Yu HH, Jang M. Classification of Vinegar Types Using Volatile Compound Profiles and Machine Learning. *Food Chemistry* 2026;514:149076. https://doi.org/10.1016/j.foodchem.2026.149076.

[77] Pourmoradian A, Barzegar M, Carbonell-Barrachina ÁA, Noguera-Artiaga L. Rapid Discrimination of Beef, Pork and Rabbit Meat Using HS-SPME Coupled with GC-MS and Chemometric Analysis. *Applied Food Research* 2026;6:102096. https://doi.org/10.1016/j.afres.2026.102096.

[78] Peng Q, Huang J, Li S, Massou BB, Chen Z, Zhu Q, et al. Comprehensive Origin Authentication of Wolfberry Pulp (Lycium Barbarum L.) Using Multimodal Sensory Analysis and Chemometrics. *Industrial Crops and Products* 2024;219:119023. https://doi.org/10.1016/j.indcrop.2024.119023.

[79] Wen J, Li J, Wang D, Li C, Robbat A, Xia L. Identification of Geographical Origin of Winter Jujube Based on GC–MS Coupled with Machine-Learning Algorithms. *Journal of Food Composition and Analysis* 2023;124:105710. https://doi.org/10.1016/j.jfca.2023.105710.

[80] Yun J, Cui C, Zhang S, Zhu J, Peng C, Cai H, et al. Use of Headspace GC/MS Combined with Chemometric Analysis to Identify the Geographic Origins of Black Tea. *Food Chemistry* 2021;360:130033. https://doi.org/10.1016/j.foodchem.2021.130033.

[81] Lakshitha R, Kantono K, Chen T, Le TT, Gannabathula S, Hamid N. Volatile Fingerprinting and Interpretable Machine Learning for Authenticating New Zealand Monofloral Honeys. *Food Research International* 2026;233:118954. https://doi.org/10.1016/j.foodres.2026.118954.

[82] Meoni G, Vita C, Tenori L, Venturini L, Ascolese M, Mattei A, et al. Machine Learning Supported Olive Oil Compound Profiling for Assessing Geographic and Cultivar Authenticity. *Food Research International* 2026;233:118890. https://doi.org/10.1016/j.foodres.2026.118890.

[83] Mamede R, Santos A, Sousa A, Díaz S, Marques SC, Leandro SM, et al. Fatty Acids Profiling of Goose Barnacle (Pollicipes Pollicipes) Tissues to Evaluate Nutritional Quality and Confirm Harvesting Location. *Journal of Food Composition and Analysis* 2024;127:105930. https://doi.org/10.1016/j.jfca.2023.105930.

[84] Yu D, Zhang X, Guo S, Yan H, Wang J, Zhou J, et al. Headspace GC/MS and Fast GC e-Nose Combined with Chemometric Analysis to Identify the Varieties and Geographical Origins of Ginger (Zingiber Officinale Roscoe). *Food Chemistry* 2022;396:133672. https://doi.org/10.1016/j.foodchem.2022.133672.

[85] Li T, Le HM, Handoyo R, Pagliano E, Hu Y. Development of Non-Targeted Volatilomics with Solid-Phase Microextraction for the Authentication of Plant-Based Milk Alternatives. *Talanta* 2025;294:128239. https://doi.org/10.1016/j.talanta.2025.128239.

[86] Aurum FS, Zaman MZ, Purwanto E, Praseptiangga D, Nakano K. Coffee Authentication via Targeted Metabolomics and Machine Learning: Unveiling Origins and Their Discriminating Biochemicals. *Food Bioscience* 2023;56:103122. https://doi.org/10.1016/j.fbio.2023.103122.

[87] Ain Syaqirah Sapian N, Aidilfitri Mohamad Roslan M, Mohd Hashim A, Nasir Mohd Desa M, Halim M, Noorzianna Abdul Manaf Y, et al. Differentiation of Lard from Other Animal Fats Based on N-Alkane Profiles Using Chemometric Analysis. *Food Research International* 2023;164:112332. https://doi.org/10.1016/j.foodres.2022.112332.

[88] Liao J, Xiao Y, Zhong J, He Y, Qian W, Chen J, et al. Origin-Dependent Chemical Profiles and Quality Differentiation of Candied Prunus Mume: Insights from Metabolomics, GC–MS Volatile Profiling, and Interpretable Machine Learning. *Food Chemistry: X* 2026;39:104391. https://doi.org/10.1016/j.fochx.2026.104391.

[89] Leite V dos SA, Ikehara BRM, Almeida NR, Silva GH, Macedo WR, Pinto FG. Metabolomics Based on GC-MS Combined with Chemometrics for Geographical Discrimination of Garlic (Allium Sativum L.). *Food Control* 2025;169:110976. https://doi.org/10.1016/j.foodcont.2024.110976.

[90] Shen C, Wang R, Jin Q, Chen X, Cai K, Xu B. Chemometrics Methods, Sensory Evaluation and Intelligent Sensory Technologies Combined with GAN-Based Integrated Deep-Learning Framework to Discriminate Salted Goose Breeds. *Food Chemistry* 2024;461:140919. https://doi.org/10.1016/j.foodchem.2024.140919.

[91] Zou Z, Wang Q, Li M, Zhen J, Yuan D, Zhang Q, et al. Combination of Gas Chromatography-Mass Spectrometry and Hyperspectral Imaging for Identification of Adulterated Safflower Seed Oil. *Journal of Food Composition and Analysis* 2024;135:106593. https://doi.org/10.1016/j.jfca.2024.106593.

[92] Torres-Cobos B, Quintanilla-Casas B, Vicario G, Guardiola F, Tres A, Vichi S. Revealing Adulterated Olive Oils by Triacylglycerol Screening Methods: Beyond the Official Method. *Food Chemistry* 2023;409:135256. https://doi.org/10.1016/j.foodchem.2022.135256.

[93] Bagnulo E, Scavarda C, Bortolini C, Cordero C, Bicchi C, Liberto E. Cocoa Quality: Chemical Relationship of Cocoa Beans and Liquors in Origin Identitation. *Food Research International* 2023;172:113199. https://doi.org/10.1016/j.foodres.2023.113199.

[94] Cecchi L, Migliorini M, Giambanelli E, Cane A, Zanoni B, Canuti V, et al. Is the Volatile Compounds Profile a Suitable Tool for Authentication of Virgin Olive Oils (Olea Europaea L.) According to Cultivars? A Study by Using HS-SPME-GC-MS and Chemometrics. *Food Control* 2022;139:109092. https://doi.org/10.1016/j.foodcont.2022.109092.

[95] Torres-Cobos B, Quintanilla-Casas B, Romero A, Ninot A, Alonso-Salces RM, Toschi TG, et al. Varietal Authentication of Virgin Olive Oil: Proving the Efficiency of Sesquiterpene Fingerprinting for Mediterranean Arbequina Oils. *Food Control* 2021;128:108200. https://doi.org/10.1016/j.foodcont.2021.108200.

[96] Quintanilla-Casas B, Bertin S, Leik K, Bustamante J, Guardiola F, Valli E, et al. Profiling versus Fingerprinting Analysis of Sesquiterpene Hydrocarbons for the Geographical Authentication of Extra Virgin Olive Oils. *Food Chemistry* 2020;307:125556. https://doi.org/10.1016/j.foodchem.2019.125556.

[97] Quintanilla-Casas B, Torres-Cobos B, Guardiola F, Servili M, Alonso-Salces RM, Valli E, et al. Geographical Authentication of Virgin Olive Oil by GC–MS Sesquiterpene Hydrocarbon Fingerprint: Verifying EU and Single Country Label-Declaration. *Food Chemistry* 2022;378:132104. https://doi.org/10.1016/j.foodchem.2022.132104.

[98] Torres-Cobos B, Quintanilla-Casas B, Rovira M, Romero A, Guardiola F, Vichi S, et al. Prospective Exploration of Hazelnut's Unsaponifiable Fraction for Geographical and Varietal Authentication: A Comparative Study of Advanced Fingerprinting and Untargeted Profiling Techniques. *Food Chemistry* 2024;441:138294. https://doi.org/10.1016/j.foodchem.2023.138294.

[99] Quintanilla-Casas B, Torres-Cobos B, Guardiola F, Romero A, Tres A, Vichi S. Geographical Authentication of Virgin Olive Oil by GC-MS Sesquiterpene Hydrocarbon Fingerprint: Scaling down to the Verification of PDO Compliance. *Food Control* 2022;139:109055. https://doi.org/10.1016/j.foodcont.2022.109055.

[100] Cecchi L, Migliorini M, Giambanelli E, Rossetti A, Cane A, Mulinacci N, et al. Authentication of the Geographical Origin of Virgin Olive Oils from the Main Worldwide Producing Countries: A New Combination of HS-SPME-GC-MS Analysis of Volatile Compounds and Chemometrics Applied to 1217 Samples. *Food Control* 2020;112:107156. https://doi.org/10.1016/j.foodcont.2020.107156.

[101] Khakimov B, Bakhytkyzy I, Fauhl-Hassek C, Engelsen SB. Non-Volatile Molecular Composition and Discrimination of Single Grape White Wines of Chardonnay, Riesling, Sauvignon Blanc and Silvaner Using Untargeted GC-MS Analysis. *Food Chemistry* 2022;369:130878. https://doi.org/10.1016/j.foodchem.2021.130878.

[102] Van De Steene J, Ruyssinck J, Fernandez-Pierna J-A, Vandermeersch L, Maes A, Van Langenhove H, et al. Authenticity Analysis of Oregano: Development, Validation and Fitness for Use of Several Food Fingerprinting Techniques. *Food Research International* 2022;162:111962. https://doi.org/10.1016/j.foodres.2022.111962.

[103] Van De Steene J, Ruyssinck J, Fernandez-Pierna J-A, Vandermeersch L, Maes A, Van Langenhove H, et al. Fingerprinting Methods for Origin and Variety Assessment of Rice: Development, Validation and Data Fusion Experiments. *Food Control* 2023;151:109780. https://doi.org/10.1016/j.foodcont.2023.109780.