SCIENTIA AGROPECUARIA

Scientia Agropecuaria

Web page: http://revistas.unitru.edu.pe/index.php/scientiaagrop

Facultad de Ciencias Agropecuarias

Universidad Nacional de Trujillo

RESEARCH ARTICLE

Near infrared spectroscopy combined with chemometrics and supervised machine learning for detection of water adulteration in raw bovine milk

Audino Melgar M.[1] .; Stephany M. Reyes[2] ; Elvys R. Santana[1] ; Jose Troestch[1, 3] *

[1] Laboratorio de Bromatología, Estación Experimental de Gualaca, Instituto de Innovación Agropecuaria de Panamá (IDIAP), Gualaca 4067, Chiriquí, Panamá.

[2] Centro de Investigación en Recursos Naturales, Facultad de Ciencias Naturales y Exactas, Universidad Autónoma de Chiriquí, 0427 Chiriquí, Panamá.

[3] Departamento de Química, Facultad de Ciencias Naturales y Exactas, Universidad Autónoma de Chiriquí, 0427 Chiriquí, Panamá.

* Corresponding author: jose.troestch@idiap.gob.pa (J. Troestch).

Received: 18 February 2026. Accepted: 14 September 2026. Published: 24 September 2026.

Abstract

Spectroscopic methods and chemometrics/machine learning offer a fast, easy, and simple approach to detect adulteration in milk. The aim of this study was to evaluate the performance of several commonly used classification and regression algorithms for detecting milk adulteration, using added water as the adulterant, based on near-infrared (NIR) spectroscopy. Different water/milk mixtures containing 1% to 40% added water were prepared and analyzed. Supervised classification algorithms such as linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), and support vector machines discriminant analysis (SVM-DA) were applied. Subsequently, supervised regression algorithms for quantitative detection, such as partial least squares regression (PLSR) and support vector regression (SVR), were evaluated. The PLS-DA models using autoscaling (AS) and standard normal variate (SNV) showed the best performance for validation in terms of sensitivity (SEN), specificity (SPE), accuracy (ACC), and error with values of 0.953, 0.850, 0.920, 0.098, and 0.930, 0.950, 0.937, 0.060, respectively. For quantitative determination, the best performance was achieved using partial least squares regression (PLSR) combined with standard normal variate (SNV) preprocessing of the spectral data. This model exhibited a validation root mean square error ($RMSE_{CV}$) of 3.912, a prediction Root Mean Square Error ($RMSE_P$) of 2.117, a residual predictive deviation (RPD) of 6.15, and a coefficient of determination for prediction (R^2_P) of 0.972. These results demonstrate that the combination of NIR spectroscopy with chemometric and machine learning techniques, such as PLSR, is suitable for the quantification of added water in raw milk samples.

Keywords: milk; adulteration; classification; regression; algorithms; machine learning.

DOI: https://doi.org/10.17268/sci.agropecu.2026.51

Cite this article:

Melgar, A. M., Reyes, S. M., Santana, E. R., & Troestch, J. (2026). Near Infrared Spectroscopy Combined with Chemometrics and Supervised Machine Learning for Detection of Water Adulteration in Raw Bovine Milk. *Scientia Agropecuaria*, *17*(3), 765-776.

1. Introduction

Food adulteration refers to the intentional addition of materials or products of lower quality, which have some degree of similarity to the properties of the original food products. This is one of the most common forms of food fraud (Aslam et al., 2023). Although these practices have a primarily economic origin or motivation, they can have direct consequences on consumer health (Allende-Prieto et al., 2025; Hellberg et al., 2020).

The high nutritional and commercial value of bovine milk make it susceptible to fraudulent practices. The most common types of adulteration in raw milk aim to increase volume, boost protein content, alter physical and chemical properties, extend shelf life, or substitute species (Ionescu et al., 2023; Poonia et al., 2017). This practice has not only been applied to milk but also to its derivatives (Ferreira et al., 2024). Among the most used adulterants in cow's milk are water, starch, urea, detergent, melamine, and milk from sheep or goats, among others (Patil et al., 2024; Qi & Jiang, 2025) This situation is considerably more prevalent in developing countries due to factors such as insufficient health monitoring and a lack of legislation related to food safety issues (Azad & Ahmed, 2016). The addition of adulterants in milk has been reported in countries such as Pakistan, Brazil, India, and China (Chen et al., 2014; Handford et al., 2016).

Among the various adulterants, extraneous water is the most common milk. Not only does it dilute the nutritional quality of the milk, but it also has a significant impact on the deterioration of milk quality in the market. Moreover, if the added water contains any chemical or biological contaminants, there is a public health risk due to potential waterborne diseases, especially in developing countries (Ehsani et al., 2023; Handford et al., 2016). Various analytical techniques have been used for the detection of adulterants in milk, including liquid and gas chromatography with different detectors, biosensors, and spectroscopic methods, enzyme-linked immunosorbent assay (ELISA), DNA-based techniques (Behkami et al., 2019; Chi et al., 2024; Himshweta & Singh, 2023; Ivanova et al., 2019; Kourti et al., 2023; Li et al., 2023; Tian et al., 2022, 2023). Many of these analytical techniques require significantly expensive equipment, such as chromatographic methods, which are also destructive to the sample and generate waste that can be harmful. Therefore, methods employing spectroscopy (UV-Visible, Mid and Near-Infrared, FT-IR, Raman) are particularly favored in food adulteration testing due to their ease of implementation and the absence of complex sample preprocessing (Aslam et al., 2023; Biancolillo et al., 2020; Çolak, 2025; Fakayode et al., 2020). However, the analysis of data obtained from such food tests can be challenging due to the large volume of data involved. Therefore, these spectro utilizes techniques are combined with chemometrics and machine learning, which utilize statistical and mathematical methods to extract meaningful information from complex data (Boateng et al., 2022; Tirado-Kulieva et al., 2026).

Multivariate parametric analysis methods, such as principal component analysis (PCA), PLS-DA, and LDA, are commonly used for classifying adulterated food samples, while PLSR has been used for the quantification of these adulterants. However, non-parametric methods such as random forests (RF), K-nearest neighbors (K-NN), gradient-boosting decision trees, and SVM have been less frequently employed, although some studies have reported better results for similar purposes when combined with spectroscopic data. This highlights the potential of non-parametric techniques, particularly in situations where the analyzed datasets are complex and the relationship between the spectrum and the response variable is nonlinear (Calle et al., 2022; Goyal et al., 2024). Although both qualitative and quantitative detection of extraneous water adulteration in bovine milk has been reported using spectroscopic methods combined with chemometrics and machine learning, the application of preprocessing techniques and algorithms, particularly non-parametric ones, remains limited. Additionally, comparative studies on the performance of different algorithms are also scarce (Sitorus & Lapcharoensuk, 2024).

The aim of this study was to evaluate the performance of three classification algorithms for detecting extraneous water adulteration in raw cow's milk. The evaluated algorithms were PLS-DA, LDA, and SVM-DA, with autoscaling and SNV used as preprocessing methods. Additionally, we assessed the performance of two regression algorithms to identify the one that provides the best results for quantifying the degree of adulteration. Specifically, we evaluated PLSR and SVR, applying multiplicative scatter correction (MSC) and SNV preprocessing methods.

2. Methodology

2.1 Overall workflow description

Figure 1 summarizes the experimental and analytical workflow used to detect and quantify added-water adulteration in raw bovine milk. The procedure comprised six sequential stages: sample collection and refrigerated transport, sample preparation and controlled water adulteration, NIR spectral acquisition, spectral preprocessing, dataset construction and splitting, and machine learning modeling. Autoscaling (AS) and standard normal variate (SNV) were applied for classification, whereas multiplicative scatter correction (MSC) and SNV were used for regression. The dataset was divided into calibration (80%) and prediction (20%) sets, with Venetian blinds cross-validation applied to the calibration set. Classification was performed using LDA, PLS-DA, and SVM-DA, while PLSR and SVR were used for quantitative prediction. Detailed procedures for each stage are described in the following subsections.

2.2 Sample collection and preparation

The procedure applied for sample collection and preparation was according to that reported by Ehsani et al. (2023), with some modifications. Samples of pure raw milk were collected at Gualaca Experimental Research Station of the Agricultural Innovation Institute of Panama (IDIAP, by its Spanish acronym) in Gualaca, Chiriquí, Panama. A total of 24 dual-purpose crossbred cows (*Bos taurus* × *Bos indicus*), mean age 7.6 years, were selected. Representative samples from each cow were collected in sterile containers and transported under refrigerated conditions to the laboratory.

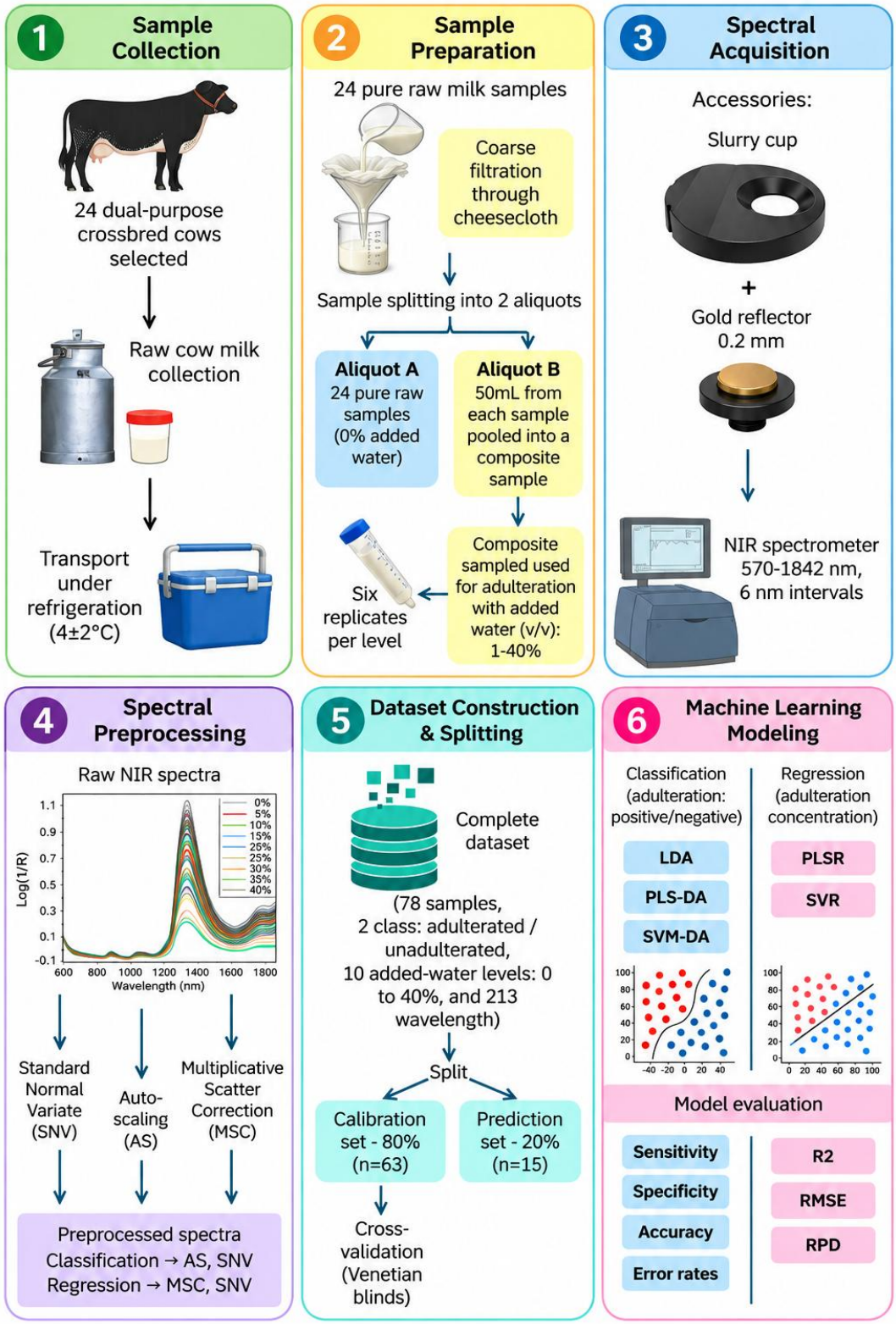

Figure 1. Workflow of the experimental design and data-analysis pipeline for the qualitative detection and quantitative determination of added-water adulteration in raw bovine milk using NIR spectroscopy and supervised machine learning.

Upon arrival at the laboratory, the raw milk samples were coarsely filtered through cheesecloth to remove coarse extraneous materials that may have been introduced during the milking process. This step was intended to improve sample homogeneity and prevent interference during NIR spectral

acquisition without altering the physicochemical composition of the milk. The use of cheesecloth provided coarse filtration only, allowing the natural milk constituents, including fat globules and dissolved components, to remain in the sample while removing only visible contaminants. Each raw milk sample was divided into two aliquots. The first aliquot was used for NIR spectral acquisition of the individual pure raw milk samples (0% added water). The second aliquot consisted of 50 mL from each of the 24 samples, which were pooled to obtain a composite milk sample. This combined reference sample was used to prepare the adulterated samples by adding tap water (pH = 6.84, conductivity = 64.00 µS/cm) at nine concentration levels, ranging from 1% to 40% (v/v). The prepared samples were manually homogenized by gentle hand shaking for 30 s using inversion movements to obtain homogeneous mixtures, after which they were subjected to spectral analysis. Six replicates were prepared for each concentration level, and the milk sample spectra were acquired at room temperature (25 ± 3 °C).

2.3 NIR spectrum acquisition

Near-infrared spectral data were acquired using an InfraXact™ Pro NIR spectrophotometer (FOSS, Denmark). Measurements were recorded over the spectral range of 570 – 1842 nm at 6 nm intervals in transflectance mode, yielding reflectance measurements at 213 discrete wavelengths (spectral variables), which were subsequently used as predictor variables for chemometric and machine learning model development. Milk samples were analyzed using the original slurry cup accessory for liquid samples (Part No. 10015160, FOSS, Denmark) together with the original gold reflector (Part No. 10013157, FOSS, Denmark) with a 0.2 mm optical path length. During spectral acquisition, the rotating slurry cup automatically collected spectra from multiple positions across the sample and averaged the individual scans to generate a representative spectrum while minimizing the effects of sample heterogeneity. The gold reflector provided stable optical reference and consistent reflectance conditions, thereby improving the repeatability and reproducibility of the spectral measurements.

2.4 Chemometric analysis

2.4.1 Datasets and preprocessing

The spectral data were arranged into a matrix of dimensions 78 × 213 and then randomly divided into a training set (80%) and a prediction set (20%). Commonly used preprocessing techniques, such as autoscaling, MSC, and SNV, were applied. A cross-validation technique (Venetian blinds, 10 data splits, 1 sample per blind) was used to evaluate the validity of the developed models.

2.4.2 Construction of classification and regression models

The dataset was categorized into two distinct classes: 24 pure (non-adulterated) samples and 54 adulterated samples. Classification algorithms included both parametric (PLS-DA, LDA) and non-parametric (SVM-DA) methods. Model performance was assessed using key metrics: sensitivity (SEN), specificity (SPE), accuracy (ACC), and their corresponding error rates to ensure model reliability and predictive power. To develop the regression models, both parametric (PLSR) and non-parametric (SVR) approaches were applied. Model performance was evaluated using regression coefficients for calibration (R^2_{CAL}), validation (R^2_{CV}), and prediction (R^2_P). Accuracy was further determined by calculating the root mean square errors for calibration ($RMSE_C$), cross-validation ($RMSE_{CV}$), and prediction ($RMSE_P$). In addition, the RPD was used as a complementary parameter to evaluate the predictive performance of the models. Data processing, figure construction, and regression and classification analyses were conducted using the SOLO software (Version 9.5, Eigenvector Research, WA, USA). All computations were performed on a 64-bit Windows 11 Pro system equipped with an Intel Core i7-12700 processor and 32 GB RAM.

3. Results and discussion

Bovine milk, due to its high economic value, is particularly vulnerable to adulteration, most commonly through the addition of water to artificially increase volume. Such unregulated or unrecorded dilution practices can pose potential health risk to consumers by altering nutritional integrity and contaminant profiles. In this context, near-infrared spectroscopy (NIRS) employing transflectance, when integrated with chemometrics and machine learning approaches, offers a rapid and reliable alternative for assessing milk quality across a wide range of food products (El Ouaddari et al., 2022).

Chemometrics and machine learning algorithms were evaluated for their capacity to detect the presence of adulterants in food and to predict the degree of adulteration (i.e., the percentage by volume of extraneous water added to a given milk sample).

The obtained NIR spectra ranged from 570 to 1842 nm in 6 nm intervals. Figure 2 illustrates the spectral data, including the average of all spectra (a),

samples classified as non-adulterated milk (green) and adulterated milk (red) (b), and samples categorized according to the percentage of added water (c). A broad, prominent band appears around 1450 nm, primarily resulting from the combined of symmetric and antisymmetric OH stretching modes of water. Notably, the intensity of the water bands increases with the added water content in the milk samples (Figure 2c) (Kasemsumran et al., 2007).

Classification models

Subsequently, the classification dataset was divided into training and validation subsets, as outlined in the data collection and analysis section. A classification algorithm was applied to differentiate adulterated samples from non-adulterated samples using the training data. Partial Least Squares Discriminant Analysis, LDA, and SVM-DA were selected for their straightforward implementation.

Linear Discriminant Analysis is a widely used machine learning algorithm used for pattern recognition. It identifies a linear combination of features that maximizes the ratio of between class variance to within class variance, thereby effectively separating two or more classes of objects (Mohammadi et al., 2024). Partial Least Squares Discriminant Analysis applies dimensionality reduction to construct a new feature space optimized for distinguishing categories, which is then used for classification and prediction (Chen et al., 2024). Support Vector Machine is a supervised learning algorithm commonly employed for binary classification and regression tasks. Its performance depends on careful tuning of parameters such as regularization constant (C), kernel type, degree, and gamma, depending on the application (Sarang, 2023). These algorithms have been successfully applied to detect a range of adulterants across diverse food matrices (Othman et al., 2023).

Figure 3 shows the predicted probability plots for milk adulteration detection using PLSDA, LDA, and SVM algorithms. Data from the prediction set was used to assess model performance and identify potential overfitting. It can be observed that SNV-PLSDA (Figure 3b) and SNV-SVM (Figure 3f) exhibited poorer performance relative to the other models, showing reduced separation between predicted probabilities for the calibration and validation sets.

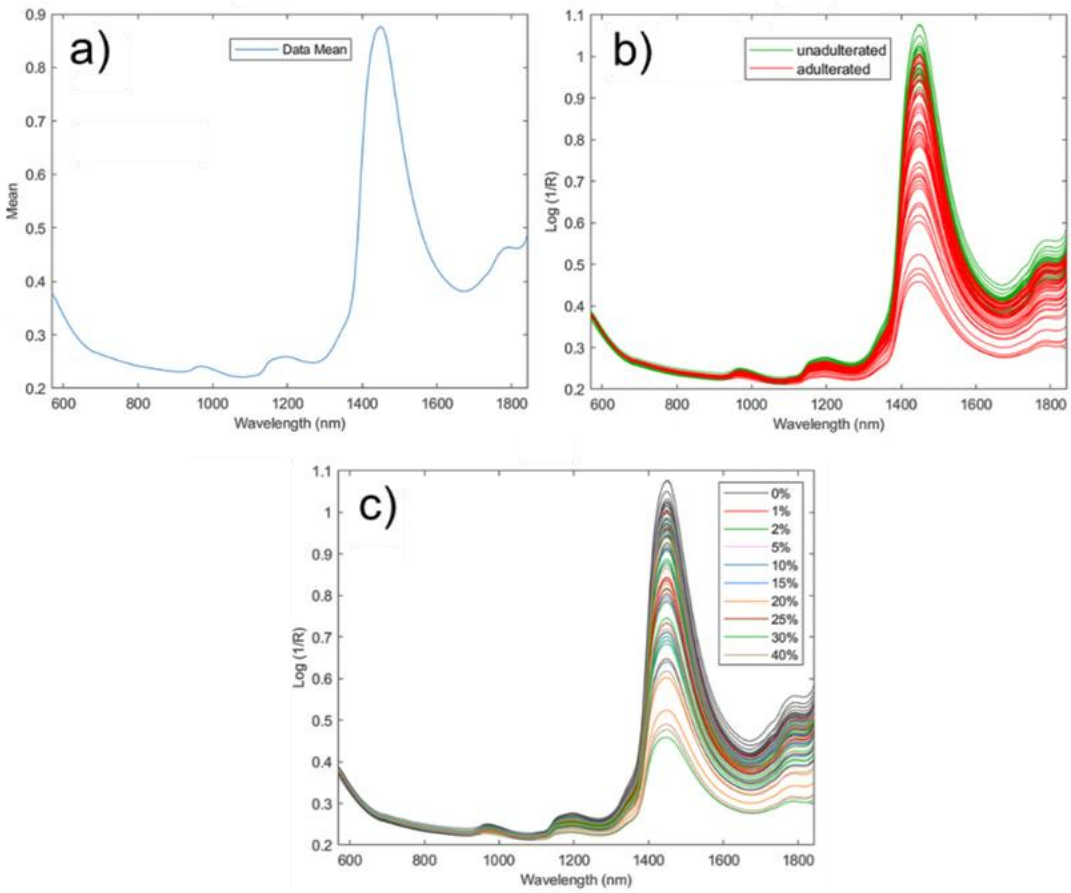

Figure 2. Near-infrared (NIR) spectral data: (a) mean of raw spectra, (b) spectra of adulterated vs. unadulterated samples, (c) spectra of pure raw milk and milk containing varying percentages of added water.

Both subsets were selected in a balanced manner to ensure that all water percentage levels were represented. Table 1 summarizes the accuracy achieved by each model for the calibration, validation, and prediction sets. Overall, model performance was higher with the training data compared to the validation data. The greater number of adulterated samples relative to pure samples lead to a higher true positive rate and, consequently, greater sensitivity compared to the true negative rate (specificity), as observed across all models.

In terms of accuracy, both LDA models yielded the lowest values during validation; specificity was also lowest for the AS-LDA model. Previous studies have indicated that when the number of variables substantially exceeds the number of samples, the performance of LDA models can be negatively affected (Boateng et al., 2022). In contrast, both PLS-DA models demonstrated the highest specificity values (0.85 and 0.95). The SNV-PLSDA model achieved the highest validation accuracy (0.937).

For the prediction set, both PLS-DA models, AS-LDA, and SNV-SVMC achieved the highest sensitivity, specificity, and accuracy values. However, among these, the AS-LDA model exhibited the highest prediction error, while the SNV-SVMC model showed lower specificity and higher error during calibration. Overall, the PLS-DA models stand out. The SNV-PLS-DA model may be preferred due to its superior validation performance (SPE = 0.950 and ACC = 0.937), which suggest greater robustness with new data. Nevertheless, the probability plot (Figure 3) indicates that the AS-PLSDA model (Figure 3a) displayed a more consistent distribution of predicted probabilities for correctly identifying adulterated samples. The superior performance of PLS-DA models compared to SVM models with other preprocessing methods for the qualitative detection of water adulteration in milk has also been reported by Ehsani et al. (2023).

Quantitative or regression models

Once the models for detecting adulteration (extraneous water) in raw milk were developed, the suitability of the technique for quantifying the percentage of adulteration was assessed. Regression algorithms were applied to all the samples generated during the adulteration process. The same dataset used for constructing the classification models was employed, with the two previously defined subsets: 80% of the samples for calibration and 20% for performance evaluation. Balance was maintained to ensure that the prediction set included at least one sample for each adulteration level.

Once the significant variables were identified, regression models were constructed using the training set. Both non-parametric (SVR) and parametric (PLSR) approaches were applied. Table 2 summarizes the key performance metrics for the models.

Table 1
Quality parameters of the optimal classification model for each algorithm

Model	Hyperparameters	Calibration				Validation				Prediction			
		SEN	SPE	ACC	ER	SEN	SPE	ACC	ER	SEN	SPE	ACC	ER
AS-PLS-DA	-	0.977	0.950	0.968	0.037	0.953	0.850	0.920	0.098	0.909	1.00	0.933	0.045
SNV-PLS-DA	-	0.953	0.950	0.952	0.048	0.930	0.950	0.937	0.060	0.909	1.00	0.933	0.045
AS-LDA	-	1.00	0.950	0.984	0.016	0.953	0.750	0.889	0.111	0.909	1.00	0.933	0.067
SNV-LDA	-	0.977	1.000	0.984	0.016	0.930	0.800	0.889	0.111	0.818	1.00	0.867	0.133
AS-SVM-DA	Cost = 10 Gamma = 0.0316	1.000	0.950	0.984	0.025	0.977	0.800	0.921	0.112	0.909	0.75	0.867	0.170
SNV-SVM-DA	Cost = 31.62 Gamma = 1.0	0.977	0.800	0.921	0.112	0.977	0.800	0.921	0.112	0.909	1.00	0.933	0.045

AS = autoscaling, SNV = standard normal variate.

Table 2
Performance of models using different modeling approaches

Model	Hyperparameters	Calibration		Validation		Prediction		RPD
		R^2_C	$RMSE_C$	R^2_{CV}	$RMSE_{CV}$	R^2_P	$RMSE_P$	
MSC-SVM	cost = 100 gamma = 10	0.909	3.969	0.846	5.144	0.876	4.789	2.72
SNV-SVM	cost = 100 gamma = 0.31623	0.915	3.847	0.845	5.152	0.887	4.644	2.80
MSC-PLSR	9 LVs	0.941	3.179	0.911	3.911	0.972	2.130	6.12
SNV-PLSR	9 LVs	0.941	3.182	0.911	3.912	0.972	2.117	6.15

MSC= Multiplicative Scatter Correction, SNV = standard normal variate.

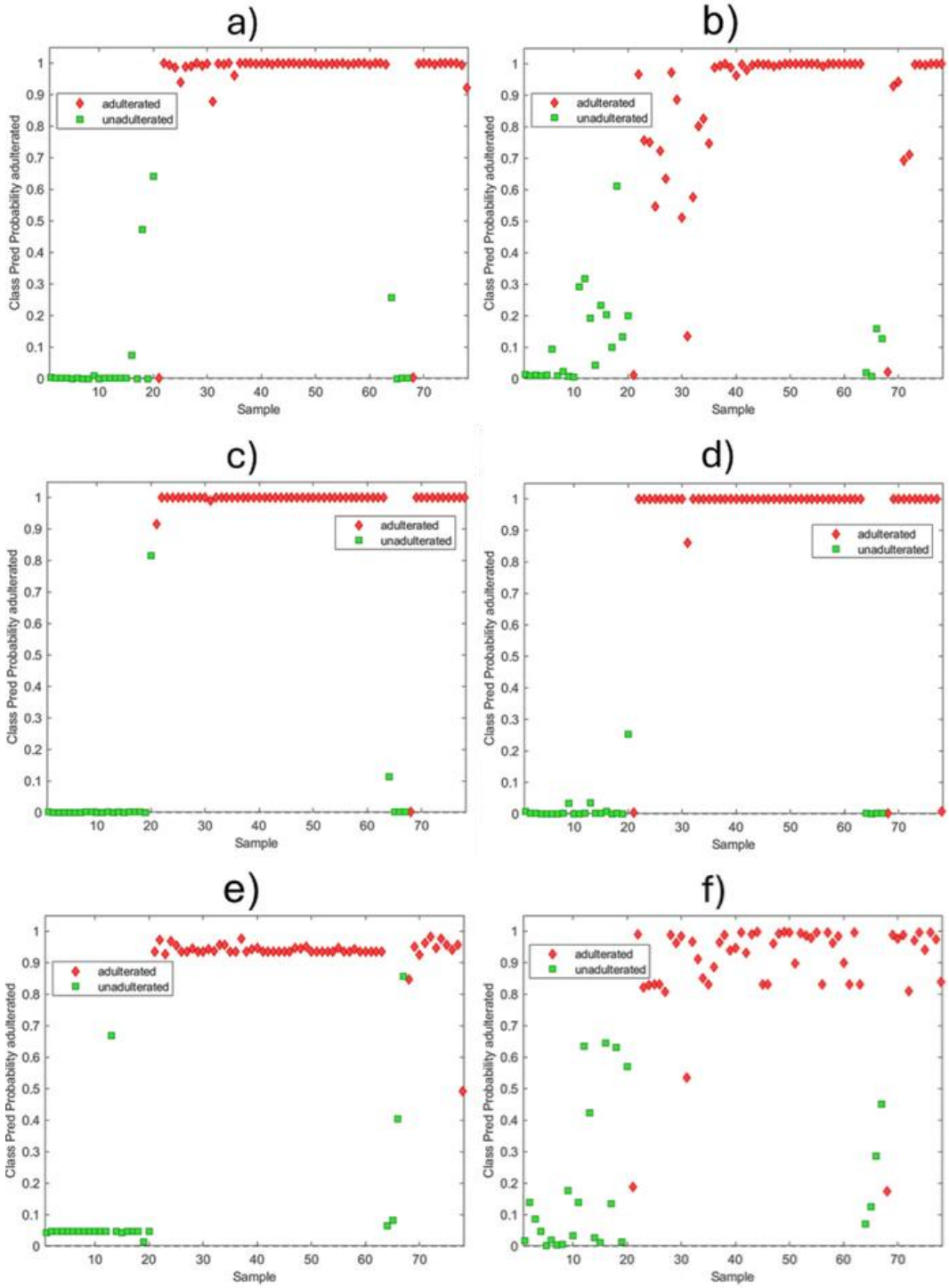

Figure 3. Y-predicted plots for milk adulteration detection using classification algorithms evaluated with different preprocessing methods: a) AS-PLSDA, b) SNV-PLSDA, c) AS-LDA, d) SNV-LDA, e) AS-SVM, and f) SNV-SVM.

For PLSR, the optimal number of components (ranging from 1 to 20) was determined through cross-validation by analyzing the evolution of the root mean square error (RMSE) as a function of component number. The best performance was achieved with nine latent variables, which yielded the greatest reduction in RMSE. The MSC-PLSR model produced a coefficient of determination (R^2) of 0.911 and an RMSE of 3.911 for the training set, and an R^2 of 0.972 with an RMSE of 2.130 for the test set. Similarly, the SNV-PLSR model achieved an R^2 of 0.911 and an RMSE of 3.912 for the training set,

and an R^2 of 0.972 with an RMSE of 2.117 for the test set, with an RPD value of 6.15. According to commonly accepted criteria, RPD values between 2.0–2.5 indicate adequate quantitative predictive capability, whereas values above 2.5 reflect excellent prediction performance (Vega et al., 2025). Although the high R^2 values demonstrate a strong correlation between spectral and response variables, the RMSE values remain relatively high for highly precise prediction of adulteration levels. Nevertheless, the coefficients of determination obtained for the PLSR models are consistent with those reported in previous studies, while differences in RMSE are expected due to variations in experimental design and sample characteristics (Ehsani et al., 2023; Mohammed & Shuming, 2021). Studies applying non-parametric regression algorithms in combination with NIR spectroscopy to quantify water adulteration in bovine milk remain scarce. Algorithms employed for this purpose include KNN, decision trees (DT), RF, and SVR (Ehsani et al., 2023; Lanjewar et al., 2024; Liang et al., 2025).

The optimal hyperparameters settings for the developed MSC-SVR and SNV-SVR models, cost = 100 and gamma = 10, and cost = 100 and gamma = 0.316, respectively (Table 2). The SNV-SVR model achieved coefficients of determination (R^2) of 0.845 and 0.887 for the training and test sets, respectively, with corresponding RMSE values of 5.152 and 4.644. The MSC-SVR model yielded R^2 values of 0.846 and 0.876 for the training and prediction sets, respectively, with RMSE values of 5.144 and 4.789. These coefficients of determination were comparable to those reported by Ehsani et al. (2023), who also employed SNV-preprocessed models, whereas higher R^2 values were reported by Liang et al. (2025). Comparison between observed and predicted values for the four regression models was conducted using scatter plots (Figure 4), which display calibration (red) and prediction (gray) datasets.

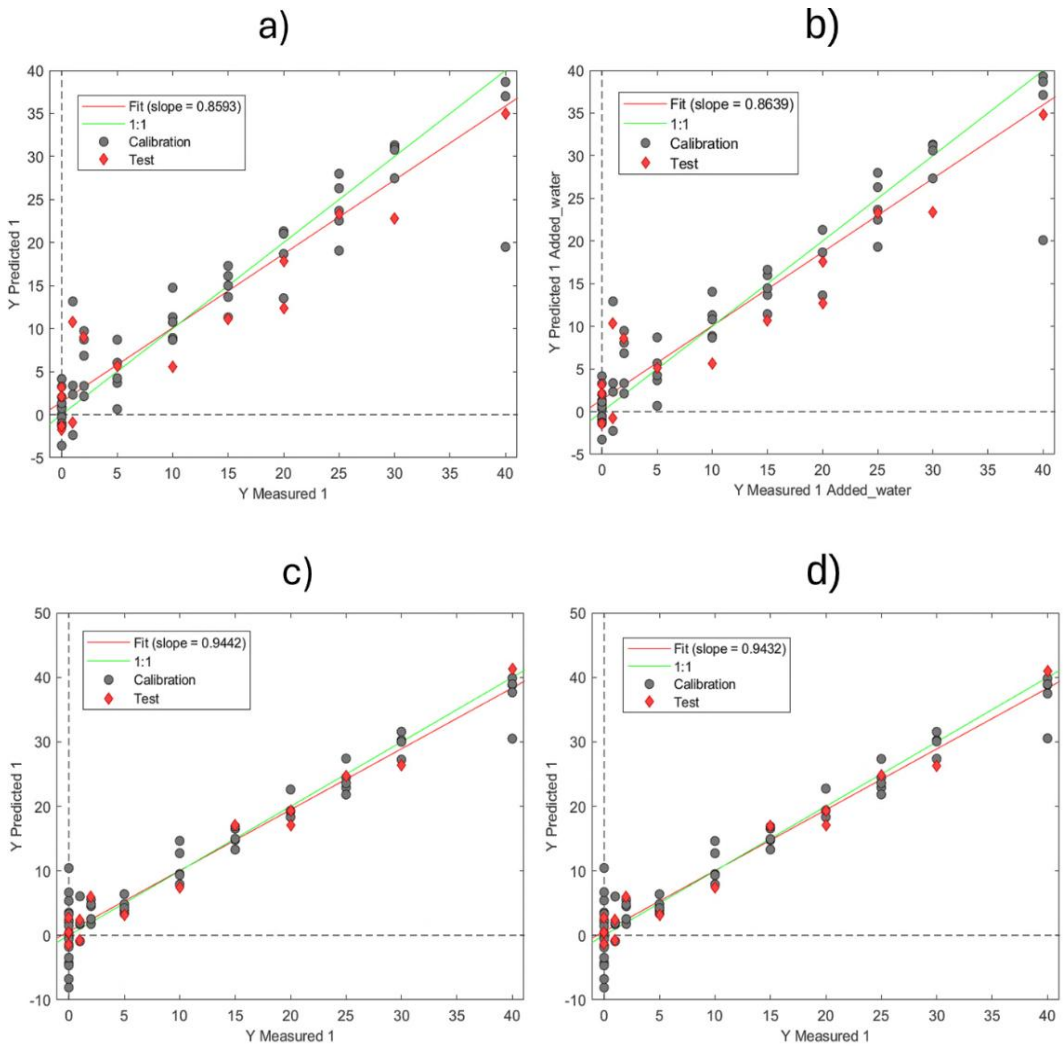

Figure 4. Prediction Plots of the Regression models (a) MSC-SVR, (b) SNV-SVR, (c) MSC-PLSR, and (d) SNV-PLSR.

In all cases, a strong correlation was observed between the reference and predicted adulteration values, with all coefficients of determination exceeding 0.845. According to the results presented in Figure 4, near-infrared spectral data exhibit a robust relationship with the water adulteration rate in milk, as demonstrated by the PLSR and SVR models. These findings highlight that technique is a valuable tool for detecting this and other forms of adulteration in bovine milk. Future research could investigate alternative preprocessing methods and algorithms to further enhance performance.

Several studies have applied NIR spectroscopy combined with chemometric and machine learning approaches for the detection and quantification of added water in bovine milk. These studies differ considerably in spectral range, preprocessing strategies, modeling algorithms, adulteration levels, and validation procedures, making direct comparison of performance metrics challenging. Nevertheless, comparison with recent reports provides a useful context for assessing the predictive capability and methodological characteristics of the models developed in the present study. Table 3 summarizes representative NIR-based studies published in recent years, highlighting their experimental ranges, spectral preprocessing methods, classification and regression algorithms, and best reported qualitative and quantitative performances.

Table 3
Comparison of recent (2022-2026) NIR-based approaches for the qualitative detection and quantitative prediction of added-water adulteration in bovine milk

Reference	Added-water range (%)	NIR spectral range (nm)	Spectral preprocessing		Algorithms		Best performance	
			Qualitative	Quantitative	Qualitative	Quantitative	Qualitative	Quantitative
(Musa, 2022)	0–50	870–1660	Not explicitly specified	Not explicitly specified	DD-SIMCA	PLSR	All test samples were reported as correctly classified; overall ACC was not explicitly reported.	PLSR: R^2val = 0.94; R^2_P = 0.93; SE_{CV} = 4.25; SE_P = 4.35
(Ehsani et al., 2023)	1–30	900–1700	Raw, EPO; SG-1st derivative + SNV also evaluated	None (BRT); SNV (PLSR)	RSDE, PLS-DA, RBF-SVM	BRT, PLSR	RSDE: ACC = 98%; Rel = 98%	BRT: R^2_{CV} = 0.95; $RMSE_{CV}$ = 0.64; R^2_P = 0.95; $RMSE_P$ = 0.58
(Lanjewar et al., 2024)	0–40	900–1700	Raw, SG, MSC, SNV + feature selection + PCA	Raw, SG, MSC, SNV + feature selection + PCA	DTC, RFC, KNNC, SVMC, LRC	DT, RF, KNN, SVR	RF: ACC = 100%; MCC = 100%	SG-PCA-KNN: R^2_V = 0.999; $RMSE_V$ = 0.399; SE_P = 0.096; RPD = 33.005
(Liang et al., 2025)	0–53.4	900–1800	None (raw spectra)	SGCS, SGFD, MSC, SNV, VN, MMN	SIMCA, NB, KNN, SVM	1D-CNN, PLSR, SVR	SIMCA: ACC = 1.000; SEN = 1.000; SPE = 1.000; PRE = 1.000; F1 = 1.000	VN-SVR: R^2_P = 0.9992; $RMSE_P$ = 0.066%; Independent validation: R^2_P = 0.9760; $RMSE_P$ = 0.2554%
Present study	**0–40**	**570–1842**	**AS, SNV**	**MSC, SNV**	**LDA, PLS-DA, SVM-DA**	**PLSR, SVR**	**SNV-PLS-DA: CV ACC = 0.937, SEN = 0.930, SPE = 0.950; Prediction ACC = 0.933, SEN = 0.909, SPE = 1.000**	**SNV-PLSR: R^2_{CV} = 0.911; $RMSE_{CV}$ = 3.912; R^2_P = 0.972; $RMSE_P$ = 2.117; RPD = 6.15**

Abbreviations: ACC: accuracy; AS: autoscaling; BRT: boosted regression tree; DD-SIMCA: data-driven soft independent modeling of class analogy; DT: decision tree; EPO: external parameter orthogonalization; F1: F1-score; KNN: k-nearest neighbors; MCC: Matthews correlation coefficient; MMN: min-max normalization; MSC: multiplicative scatter correction; NB: naive Bayes; PCA: principal component analysis; PLS-DA: partial least squares discriminant analysis; PLSR: partial least squares regression; PRE: precision; Rel: reliability; RF: random forest; RPD: residual predictive deviation; RSDE: random subspace discriminant ensemble; SG: Savitzky–Golay; SGCS: Savitzky–Golay convolutional smoothing; SGFD: Savitzky–Golay filtered derivative; SEN: sensitivity; SNV: standard normal variate; SPE: specificity; SVM: support vector machine; SVR: support vector regression; VN: vector normalization.

As shown in Table 3, the present study differs from previous NIR-based approaches in several methodological aspects. First, spectra were acquired over a broader wavelength range (570–1842 nm) than those used by Musa (2022), Ehsani et al. (2023), Lanjewar et al. (2024), and Liang et al. (2025), whose measurements were mainly restricted to the 870/900–1660/1800 nm region. Second, rather than relying on a single modeling strategy, ensemble methods, extensive feature-selection procedures, or deep-learning architectures, the present study systematically compared both parametric and non-parametric approaches for the two analytical objectives: LDA, PLS-DA, and SVM-DA for qualitative detection, and PLSR and SVR for quantitative prediction. In particular, the SNV-PLSR model demonstrated strong predictive capability using full spectral information and a relatively simple preprocessing and modeling workflow. Although lower prediction errors have been reported for some previous approaches, particularly those incorporating feature selection or more complex nonlinear algorithms, differences in sample composition, adulteration design, dataset size, and validation strategy preclude direct ranking of model performance across studies. An additional finding of the present study is that the linear latent-variable PLSR approach clearly outperformed the nonlinear SVR models under the experimental conditions evaluated, suggesting that the relationship between NIR spectral variation and added-water concentration was adequately captured by the latent linear structure of the data. Overall, these results demonstrate that high predictive performance can be achieved without requiring complex feature-selection or ensemble/deep-learning strategies, while simultaneously providing qualitative detection and quantitative estimation of added-water adulteration in raw bovine milk.

4. Conclusions

NIR spectroscopy combined with supervised machine learning demonstrated a strong ability to detect and quantify added-water adulteration in raw bovine milk under the conditions evaluated in this study. Among the classification approaches, PLS-DA provided the best overall discrimination between adulterated and non-adulterated samples, achieving a prediction accuracy of 0.933, sensitivity of 0.909, and specificity of 1.000. For quantitative prediction, SNV-PLSR showed the best performance, with R^2_P = 0.972, $RMSE_P$ = 2.117, and RPD = 6.15, substantially outperforming the corresponding SVR models. These results indicate that the spectral changes associated with increasing levels of added water can be effectively captured using a comparatively simple latent-variable model, supporting the potential of NIR spectroscopy as a rapid and non-destructive screening tool for milk authenticity assessment.

Future studies should focus primarily on evaluating model robustness and transferability using independent raw milk samples collected from different farms, production systems, seasons, and sampling periods, with particular emphasis on low adulteration levels where discrimination may be more challenging. External validation using newly collected milk batches and independently prepared adulterated samples would provide a more rigorous assessment of predictive performance. In addition, nonlinear and ensemble approaches such as Random Forest, Extreme Gradient Boosting (XGBoost), and one-dimensional convolutional neural networks (1D-CNN) could be evaluated and compared with PLS-based models.

Variable-selection strategies such as competitive adaptive reweighted sampling (CARS) or successive projections algorithm (SPA) could also be investigated to identify the most informative spectral regions and potentially simplify future models.

Acknowledgments

This study was supported by the Agricultural Innovation Institute of Panama (IDIAP), through project 501.D.2.11, and by the National Research System of Panama (SNI-AIP) of the National Secretariat of Science, Technology and Innovation of Panama (SENACYT).

Author Contributions

A. Melgar: review, editing, supervision, resources and project administration. **S. Reyes:** methodology, review and data analysis. **E. Santana:** methodology and resources. **J. Troestch:** conceptualization, investigation, data collection and analysis, writing & software.

Conflict of Interest

The authors declare no conflict of interest regarding the publication of this article.

ORCID

A. Melgar M. https://orcid.org/0000-0001-8491-0236
S. M. Reyes https://orcid.org/0000-0003-2619-6497
E. R. Santana https://orcid.org/0009-0009-2385-1194
J. Troestch https://orcid.org/0000-0003-2900-741X

References

Allende-Prieto, C., Fernández, L., Rodríguez-Gonzálvez, P., Martínez, B., García, P., & Rodríguez, A. (2025). Portable optical instrument for detection and quantification of milk adulteration: A study on mixtures from different species and water dilution. *International Dairy Journal, 164*. https://doi.org/10.1016/j.idairyj.2025.106186

Aslam, R., Sharma, S. R., Kaur, J., Panayampadan, A. S., & Dar, O. I. (2023). A systematic account of food adulteration and

recent trends in the non-destructive analysis of food fraud detection. *Journal of Food Measurement and Characterization, 17*(3), 3094–3114. https://doi.org/10.1007/s11694-023-01846-3

Azad, T., & Ahmed, S. (2016). Common milk adulteration and their detection techniques. *International Journal of Food Contamination, 3*(1). https://doi.org/10.1186/s40550-016-0045-3

Behkami, S., Zain, S. M., Gholami, M., & Khir, M. F. A. (2019). Classification of cow milk using artificial neural network developed from the spectral data of single- and three-detector spectrophotometers. *Food Chemistry*, *294*, 309–315. https://doi.org/10.1016/j.foodchem.2019.05.060

Biancolillo, A., Marini, F., Ruckebusch, C., & Vitale, R. (2020). Chemometric Strategies for Spectroscopy-Based Food Authentication. *Applied Sciences*, *10*(18), 6544. https://doi.org/10.3390/app10186544

Boateng, A. A., Sumaila, S., Lartey, M., Oppong, M. B., Opuni, K. F. M., & Adutwum, L. A. (2022). Evaluation of chemometric classification and regression models for the detection of syrup adulteration in honey. *LWT*, *163*. https://doi.org/10.1016/j.lwt.2022.113498

Calle, J. L. P., Ferreiro-González, M., Ruiz-Rodríguez, A., Fernández, D., & Palma, M. (2022). Detection of Adulterations in Fruit Juices Using Machine Learning Methods over FT-IR Spectroscopic Data. *Agronomy*, *12*(3), 1–14. https://doi.org/10.3390/agronomy12030683

Chen, C., Zhang, J., & Delaurentis, T. (2014). Quality control in food supply chain management: An analytical model and case study of the adulterated milk incident in China. *International Journal of Production Economics*, *152*, 188–199. https://doi.org/10.1016/j.ijpe.2013.12.016

Chen, Y., Li, S., Jia, J., Sun, C., Cui, E., Xu, Y., Shi, F., & Tang, A. (2024). FT-NIR combined with machine learning was used to rapidly detect the adulteration of pericarpium citri reticulatae (chenpi) and predict the adulteration concentration. *Food Chemistry: X*, *24*. https://doi.org/10.1016/j.fochx.2024.101798

Chi, S.-X., Liu, B.-H., Zhang, B., Wang, B.-R., Zhou, J., Li, L., Zhang, Y.-H., & Mu, Z. (2024). Development of an ELISA method to determine adulterated cow milk in camel milk. *International Dairy Journal*, *155*, 105953. https://doi.org/10.1016/j.idairyj.2024.105953

Çolak, S. (2025). Simultaneous Raman and FTIR-ATR Spectroscopy Techniques Combined with Chemometrics: Characterization and Comparison of Donkey Milk Adulteration. *Journal of Raman Spectroscopy*, *56*(7), 598–608. https://doi.org/10.1002/jrs.6812

Ehsani, S., Dastgerdy, E. M., Yazdanpanah, H., & Parastar, H. (2023). Ensemble classification and regression techniques combined with portable near infrared spectroscopy for facile and rapid detection of water adulteration in bovine raw milk. *Journal of Chemometrics*, *37*(1). https://doi.org/10.1002/cem.3395

El Ouaddari, A., El Amrani, A., Jamal Eddine, J., & Antonio Cayuela-Sánchez, J. (2022). Rapid prediction of essential oils major components by Vis/NIRS models using compositional methods. *Results in Chemistry*, *4*. https://doi.org/10.1016/j.rechem.2022.100562

Fakayode, S. O., Baker, G. A., Bwambok, D. K., Bhawawet, N., Elzey, B., Siraj, N., Macchi, S., Pollard, D. A., Perez, R. L., Duncan, A. V., & Warner, I. M. (2020). Molecular (Raman, NIR, and FTIR) spectroscopy and multivariate analysis in consumable products analysis. *Applied Spectroscopy Reviews*, *55*(8), 647–723. https://doi.org/10.1080/05704928.2019.1631176

Ferreira, M. M., Marins-Gonçalves, L., & De Souza, D. (2024). An integrative review of analytical techniques used in food authentication: A detailed description for milk and dairy products. *Food Chemistry*, *457*. https://doi.org/10.1016/j.foodchem.2024.140206

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

Handford, C. E., Campbell, K., & Elliott, C. T. (2016). Impacts of Milk Fraud on Food Safety and Nutrition with Special Emphasis on Developing Countries. *Comprehensive Reviews in Food Science and Food Safety*, *15*(1), 130–142. https://doi.org/10.1111/1541-4337.12181

Hellberg, R. S., Sklare, S. A., & Everstine, K. (2020). Food Fraud: A Global Threat with Public Health and Economic Consequences. In *Food Fraud: A Global Threat with Public Health and Economic Consequences*. Elsevier. https://doi.org/10.1016/B978-0-12-817242-1.00020-8

Himshweta, & Singh, M. (2023). Nanosensor platforms for detection of milk adulterants. *Sensors and Actuators Reports*, *5*, 100159. https://doi.org/10.1016/j.snr.2023.100159

Ionescu, A. D., Cîrîc, A. I., & Begea, M. (2023). A Review of Milk Frauds and Adulterations from a Technological Perspective. *Applied Sciences*, *13*(17). https://doi.org/10.3390/app13179821

Ivanova, A. S., Merkuleva, A. D., Andreev, S. V., & Sakharov, K. A. (2019). Method for determination of hydrogen peroxide in adulterated milk using high performance liquid chromatography. *Food Chemistry*, *283*, 431–436. https://doi.org/10.1016/j.foodchem.2019.01.051

Kasemsumran, S., Thanapase, W., & Kiatsoonthon, A. (2007). Feasibility of near-infrared spectroscopy to detect and to quantify adulterants in cow milk. *Analytical Sciences*, *23*(7), 907–910. https://doi.org/10.2116/analsci.23.907

Kourti, D., Angelopoulou, M., Misiakos, K., Makarona, E., Economou, A., Petrou, P., & Kakabakos, S. (2023). Detection of Adulteration of Milk from Other Species with Cow Milk through an Immersible Photonic Immunosensor. *Engineering Proceedings*, *35*(1). https://doi.org/10.3390/IECB2023-14582

Lanjewar, M. G., Parab, J. S., & Kamat, R. K. (2024). Machine learning based technique to predict the water adulterant in milk using portable near infrared spectroscopy. *Journal of Food Composition and Analysis*, *131*. https://doi.org/10.1016/j.jfca.2024.106270

Li, B., Yu, M., Xu, W., Chen, L., & Han, J. (2023). Comparison of PCR Techniques in Adulteration Identification of Dairy Products. *Agriculture*,*13*(7). https://doi.org/10.3390/agriculture13071450

Liang, Q., Xia, Y. F., Che, J. K., Liu, Y., Zhang, H., Guo, J. C., Xu, Q., & Xue, H. N. (2025). Detection of water adulteration levels in milk using near-infrared spectroscopy combined with chemometrics. *Journal of Dairy Science*, *108*(7), 6852–6866. https://doi.org/10.3168/jds.2025-26631

Mohammadi, N., Esteki, M., & Simal-Gandara, J. (2024). Machine learning for authentication of black tea from narrow-geographic origins: Combination of PCA and PLS with LDA and SVM classifiers. *LWT*, *203*. https://doi.org/10.1016/j.lwt.2024.116401

Mohammed, A. M., & Shuming, Y. (2021). Detection and quantification of cow milk adulteration using portable near-infrared spectroscopy combined with chemometrics. *African Journal of Agricultural Research*, *17*(2), 198–207. https://doi.org/10.5897/ajar2020.15321

Musa, M. A. (2022). Evaluation of portable NIR assay for detection of milk adulteration with water. *Advances in Dairy Research*, *10*(1). https://doi.org/10.35248/2329-888X.22.10.599

Othman, S., Mavani, N. R., Hussain, M. A., Rahman, N. A., & Mohd Ali, J. (2023). Artificial intelligence-based techniques for adulteration and defect detections in food and agricultural industry: A review. *Journal of Agriculture and Food Research*, *12*. https://doi.org/10.1016/j.jafr.2023.100590

Patil, G. B., Wani, S. P., Bafna, P. S., Bagul, V. S., Kalaskar, M. G., & Mutha, R. E. (2024). Milk adulteration: From detection to health impact. *Food and Humanity*, *3*, 100339. https://doi.org/10.1016/j.foohum.2024.100339

Poonia, A., Jha, A., Sharma, R., Singh, H. B., Rai, A. K., & Sharma, N. (2017). Detection of adulteration in milk: A review.

International Journal of Dairy Technology, *70*(1), 23–42. https://doi.org/10.1111/1471-0307.12274

Qi, W., & Jiang, Q. (2025). Enhanced near-infrared spectroscopy for rapid and sensitive detection of melamine in milk: a novel methodological approach. *Chemical and Biological Technologies in Agriculture*, *12*(1). https://doi.org/10.1186/s40538-025-00863-2

Sarang, P. (2023). Support Vector Machines. In *Thinking Data Science* (pp. 153–165). Springer, Cham. https://doi.org/10.1007/978-3-031-02363-7_8

Sitorus, A., & Lapcharoensuk, R. (2024). Development of automatic tuning for combined preprocessing and hyperparameters of machine learning and its application to NIR spectral data of coconut milk adulteration. *Food Chemistry*, *457*, 140108. https://doi.org/10.1016/j.foodchem.2024.140108

Tian, H., Chen, S., Li, D., Lou, X., Chen, C., & Yu, H. (2022). Simultaneous detection for adulterations of maltodextrin, sodium carbonate, and whey in raw milk using Raman spectroscopy and chemometrics. *Journal of Dairy Science*, *105*(9), 7242–7252. https://doi.org/10.3168/jds.2021-21082

Tian, H., Xiong, J., Chen, S., Yu, H., Chen, C., Huang, J., Yuan, H., & Lou, X. (2023). Rapid identification of adulteration in raw bovine milk with soymilk by electronic nose and headspace-gas chromatography ion-mobility spectrometry. *Food Chemistry: X*, *18*. https://doi.org/10.1016/j.fochx.2023.100696

Tirado-Kulieva, V. A., Gonzales-Malca, J. A., Seminario-Sanz, R. S., & Castro, W. (2026). A systematic review and critical analysis of the combination of near-infrared spectroscopy and chemometrics for milk quality assessment. *Applied Food Research*, *6*(2). https://doi.org/10.1016/j.afres.2026.102372

Vega, A., Reyes, S. M., & Troestch, J. (2025). Physicochemical Parameters and Multivariate Analysis to Predict the Sensory Quality in Specialty Coffee from Panama. *ACS Omega*, *10*(13), 13251–13259. https://doi.org/10.1021/acsomega.4c10914