Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 9, S. 31-40, 6, 2025

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 9, pp. 31-40, 6, 2025

Research Article

Rapid and Accurate Estimation of Milk Fat by Near Infrared Spectroscopy: Comparison of Different Pre-processing and Regression Methods

Özcan Çataltaş*

Electrical and Electronic Engineering Department, Selcuk University, Türkiye

(ozcancataltas@selcuk.edu.tr)

(Received: 27 May 2025, Accepted: 04 June 2025)

(5th International Conference on Contemporary Academic Research ICCAR 2025, May 30-31, 2025)

ATIF/REFERENCE: Çataltaş, Ö. (2025). Rapid and Accurate Estimation of Milk Fat by Near Infrared Spectroscopy: Comparison of Different Pre-processing and Regression Methods. *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(6), 31-40.

Abstract –This study aims to rapidly and accurately estimate the fat content of milk using near-infrared spectroscopy and various chemometric analysis methods. In the study, different pre-processing techniques such as standard normal variate, multiplicative scatter correction, Savitzky-Golay smoothing, and spectral differentiation were applied along with various modeling approaches such as partial least squares regression, ridge regression, support vector regression, lasso regression, and random forest regression. The findings show that pre-processing methods have a decisive impact on model success. In particular, the use of standard normal variate and first derivative pre-processing methods in combination with partial least squares regression and ridge regression resulted in the highest accuracy and lowest error values. The results suggest that near-infrared spectroscopy can be an effective and reliable tool for automation and real-time monitoring of quality control processes in the dairy industry.

Keywords – Near-Infrared Spectroscopy, Regression, Pre-Processing, Fat Detection, Chemometry.

I. INTRODUCTION

Milk has an important place as a basic food in human nutrition. Milk, which contains many nutrients such as protein, carbohydrates, vitamins, minerals, and fat, is an indispensable food source, especially for children, the elderly, and the sick [1]. One of the most important parameters determining the nutritional value of milk is its fat content. Milk fat is of critical nutritional importance as it is a source of energy and plays a role in the transportation and absorption of fat-soluble vitamins (A, D, E, K) [2]. In addition, milk fat directly affects the flavor, aroma, and texture of milk. Therefore, accurate and rapid determination of fat content is of great importance in the quality control of milk and dairy products.

Various chemical and physical analysis methods are traditionally used to determine the fat content of milk. One of the most widely used methods is the Gerber method, in which milk fat is separated and measured volumetrically with the help of sulfuric acid and amyl alcohol added to milk. Other techniques, such as gravimetric methods and the Babcock method, are also widely used [3]. However, these traditional methods are often time-consuming, costly, and sometimes involve the use of environmentally hazardous chemicals. Furthermore, the applicability of these methods can be limited when high sample numbers are required [4].

In recent years, the need for fast, reliable, and environmentally friendly alternative methods for the analysis of milk and dairy products has increased. In this context, Near Infrared (NIR) spectroscopy has become one of the leading modern techniques in milk analysis [5], [6]. NIR spectroscopy is a method that works in the wavelength range of 780-2500 nm and can be analyzed without damaging the sample and without using any chemical reagents. The basic principle of NIR spectroscopy is that the characteristic vibrations of organic molecules show absorption in this wavelength range. In complex matrices such as milk, components such as fat, protein, and lactose have unique signals in the NIR spectrum. This enables rapid and non-invasive estimation of important parameters such as milk fat content [7].

The advantages of NIR spectroscopy in milk analysis include fast measurement time, no sample preparation, multi-parameter analysis, and no use of environmentally harmful chemicals. Furthermore, NIR instruments are portable and can be used in field applications. However, since NIR spectra often contain complex and overlapping signals, advanced data processing and calibration techniques are needed to obtain accurate and reliable results [8].

In this study, a regression analysis of milk fat content was performed using a readily available milk NIR spectroscopy dataset. The dataset was divided into calibration and validation sets using the Kennard-Stone (KS) algorithm. Then, various pre-processing methods were applied to the spectral data. The pre-processed data were analyzed using different regression models. The main objective of the study is to compare the performance of different combinations of pre-processing and regression models in the estimation of milk fat content and to determine the most appropriate approach. The findings aim to demonstrate the potential of NIR spectroscopy for fast and reliable fat determination in the dairy industry.

II. MATERIALS AND METHOD

A. Dataset Description and Split

The dataset used for the regression analysis of milk fat content in this study is a publicly available NIR spectroscopy dataset published by Jose A. Diaz-Olivares et al. [9]. The dataset contains NIR spectral measurements of 1224 raw milk samples collected over eight weeks in 2017 at the Hooibeekhoeve experimental farm in Antwerp, Belgium. Measurements were made by taking a representative sample of raw milk at each milking with an automatic milking system. Each milk sample was obtained from 41 Holstein cows with an average lactation period of 168 ± 84 days and an average number of calvings of 2.0 ± 1.1. Milk samples were collected from a total of 1270 milkings, averaging 158 per week; however, 1224 samples with complete laboratory reference analyses and spectral measurements were used in the analysis.

An NIR spectrometer (1.7-256 Plane Grating Spectrometer, Carl Zeiss, Jena, Germany) with a 256-pixel cooled InGaAs diode array operating in the wavelength range 960-1690 nm was used for each milk sample. Spectral measurements were averaged over 100 repetitions for each sample with a resolution of 2.86 nm/pixel and an integration time of 100 ms. The measurement system was equipped with a 20 W integrated halogen light source, milk flow control by a peristaltic pump, a special borosilicate cuvette, white reference, and dark reference. For each milk sample, the white and dark reference spectra were recorded along with the sample spectrum. These references were used to correct for variations in light source intensity and spectrometer sensitivity. The spectral data were normalized with the following formula:

$$Normalized\ Spectra = \frac{Sample\ Spectra - Dark\ Spectra}{White\ Spectra - Dark\ Spectra}$$

The data set contains chemical and physical parameters of each milk sample, such as cow ID, week, milk yield, fat, protein, lactose, urea, SCC, and NIR spectral data normalized at 256 wavelengths.

In the study, during the modeling process, the data set was divided into calibration (training) and validation (validation) sets using the Kennard-Stone (KS) algorithm. The Kennard-Stone algorithm is a widely used sample selection method in multivariate data analysis. Its primary purpose is to select a subset that best represents the distribution of samples in the data set. This algorithm is used to increase the

generalizability of the model, especially in the construction of calibration and validation sets [10]. The working principle of the KS algorithm is as follows:

Selection of Initial Samples: The algorithm starts by selecting the two samples with the largest distance between the samples in the dataset. These two samples are determined as the initial members of the calibration set.

Adding Samples: For the remaining samples, the smallest distances to the samples in the calibration set are calculated. Then, the largest of these smallest distances is selected. This means identifying the sample that is furthest from the calibration set. This sample is also added to the calibration set.

Iteration: The second step is repeated until the desired number of samples is added to the calibration set. At each step, the sample farthest from the calibration set is selected, increasing the diversity and representativeness of the set.

The KS algorithm ensures that the calibration and validation sets are representative and homogeneous by taking into account the distribution of the samples in the spectral space. This improves the generalizability and accuracy of the model. The validation set consists of samples that are not included in the calibration set and are used to evaluate the performance of the model independently.

B. Pre-processing Methods

Raw spectral data from NIR spectroscopy often contains scattering effects, baseline shifts, noise, and other physical/matrix-induced variations. Such unwanted variations can adversely affect the performance of modeling based on chemical information. For this reason, various pre-processing methods are applied to spectral data to improve the signal-to-noise ratio and highlight variations based on chemical information [11]. The pre-processing methods used in this study and their mathematical basis are described in detail below.

Standard Normal Variate

Standard Normal Variate (SNV) is a pre-processing method used to correct scattering effects and baseline shifts in spectral data [12]. SNV normalizes each spectrum by its mean and standard deviation. Thus, each spectrum is rescaled so that its mean is zero and its standard deviation is one. Assuming a spectrum is defined as $x = [x_1, x_2, \ldots, x_n]$, the SNV transformation is applied as follows:

$$x_i^{SNV} = \frac{x_i - \bar{x}}{s}$$

Where x_i is the value of the original spectrum at wavelength i, $\bar{x}$ is the mean of the spectrum, and s is the standard deviation of the spectrum. SNV is particularly effective in reducing scattering variations due to particle size differences or heterogeneity on the sample surface.

Multiplicative Scatter Correction

Multiplicative Scatter Correction (MSC) is a method for correcting multiplicative and additive scattering effects in spectral data. MSC linearly aligns each spectrum to the average spectrum chosen as a reference [13]. For a spectrum x and a reference spectrum x_{ref}, the MSC transformation is applied as follows:

First, linear regression is performed between x and x_{ref}:

$$x = a + b.x_{ref} + \epsilon$$

The MSC corrected spectrum is obtained as follows:

$$x_i^{MSC} = \frac{x - a}{b}$$

In particular, MSC reduces scattering variations due to physical differences in the sample matrix and emphasizes variations based on chemical composition.

Savitzky-Golay (SG) Smoothing

Savitzky-Golay (SG) filtering is a smoothing and differentiation method used to reduce noise in spectral data and preserve the fundamental structure of the signal. The SG filter applies a moving polynomial regression with a given window width and a given polynomial degree [14]. Assuming that a spectrum is denoted by x, the SG filter calculates a new value by polynomial fitting a given range around each data point. The SG filter is also commonly used to take the first and second derivatives of the spectrum.

The mathematical basis of the SG filter is to fit the following polynomial for each window:

$$y_i = \sum_{k=0}^{d} a_k (x_i)^k$$

Here, d is the degree of the polynomial, a_k is the polynomial coefficients. The SG filter reduces high-frequency noise in spectral data while preserving the main structure and peaks of the signal.

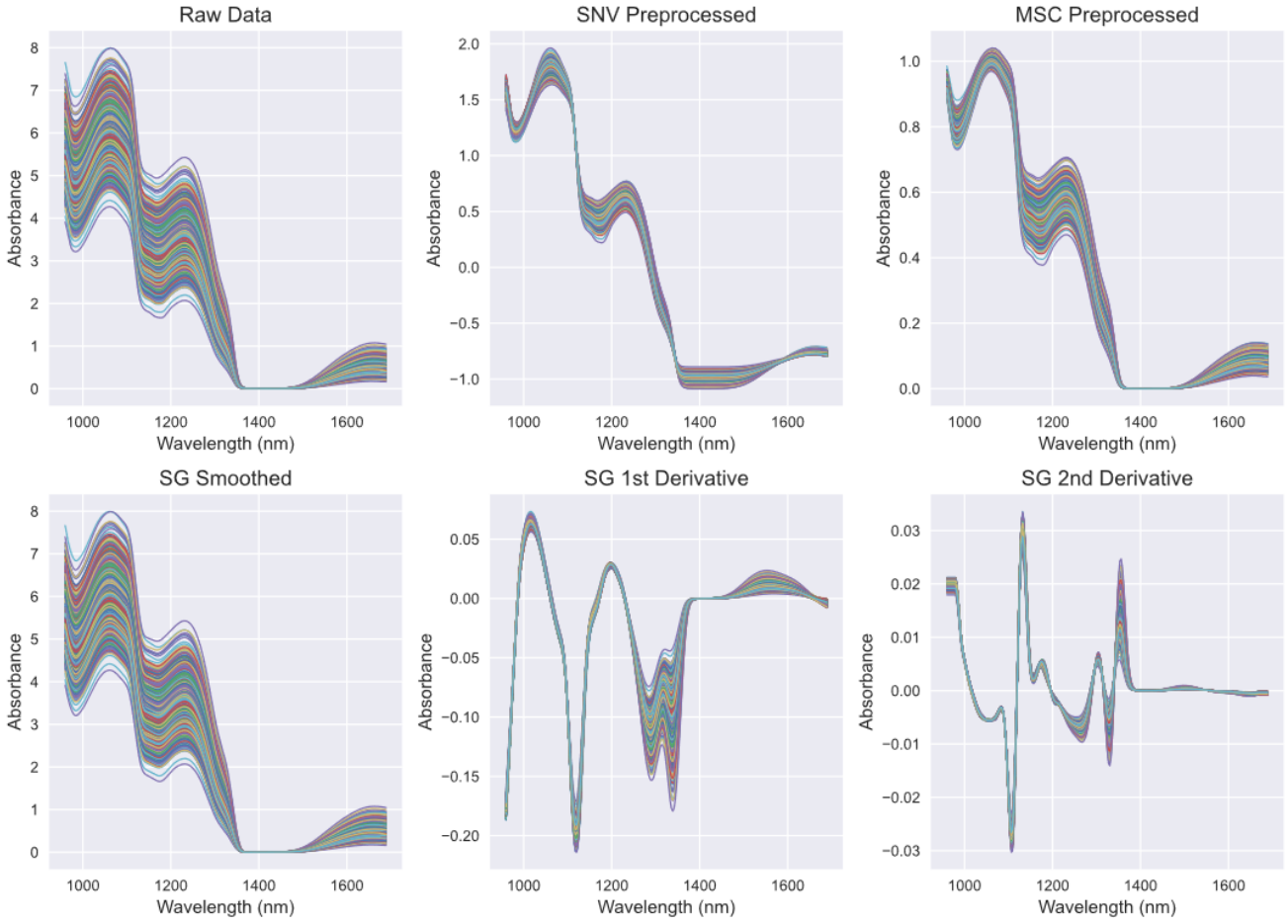

Fig. 1 The original and pre-processed spectra

First Derivative

First derivative pre-processing is used to remove baseline shifts and fixed offsets in spectral data. It also helps to separate overlapping peaks. The first derivative shows the rate of change of the spectral signal with respect to wavelength.

Second Derivative

Second derivative pre-processing is used further to reduce baseline shifts and trends in spectral data, better distinguish overlapping peaks, and reveal fine details of the signal. The second derivative shows the change in the slope of the signal. The second derivative more effectively removes baseline shifts and trends in the spectral data and facilitates the separation of overlapping peaks. However, it is usually applied in combination with the SG filter as it can increase high-frequency noise.

The original and pre-processed spectra are given in Fig. 1.

C. Regression Methods

Spectral data obtained by NIRS are generally high dimensional and multivariate. Therefore, classical linear regression methods may be inadequate for the estimation of milk fat content [15]. In this study, five different regression models with different mathematical approaches were used: Partial Least Squares Regression (PLSR), Ridge Regression, Support Vector Regression (SVR), Lasso Regression, and Random Forest regression. The theoretical foundations and parameter optimization of each model are described in detail below.

Partial Least Squares Regression

PLSR is a generalization of multiple linear regression, especially for high-dimensional and multicollinear data. PLSR creates new latent variables that maximize the common variance between both the independent variables (X: spectral data) and the dependent variables (Y: fat content) [15].

The mathematical basis of the PLSR model is as follows:

$$X = TP^T + E$$

$$Y = UQ^T + F$$

Where X is the matrix of independent variables (spectral data), Y is the matrix of dependent variables, T and U are latent score matrices, P and Q are loading matrices, and E and F are error matrices. PLSR is optimized by the latent variable parameter. This parameter is usually determined by cross-validation.

Ridge Regression

Ridge regression is a variant of linear regression that avoids overfitting by shrinking the coefficients of the model. In Ridge regression, a penalty term is added to the sum of error squares [16].

Mathematical formulation of Ridge regression:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Here, β is the regression coefficient, and λ is the penalty parameter. In Ridge regression, the penalty parameter controls the complexity of the model and is usually optimized by cross-validation.

Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression adds a penalty term similar to Ridge regression, but here, the penalty is the sum of the absolute values of the coefficients [17]. In this way, some coefficients can be set to zero, and variable selection can be made.

Mathematical formulation of Lasso regression:

$$\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

Here, λ, is the penalty parameter. In Lasso, the penalty parameter is optimized by cross-validation.

Support Vector Regression

SVR is a powerful machine learning method for linear and nonlinear regression problems. SVR uses kernel functions to transform the data into a high-dimensional space and tries to find the best fit within an error tolerance [18].

The mathematical basis of SVR:

$$\min_{\mathbf{w},b,\xi,\xi^*}\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$y_i - (\mathbf{w}^T\phi(x_i) + b) \le \epsilon + \xi_i$$
$$(\mathbf{w}^T\phi(x_i) + b) - y_i \le \epsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0$$

Where w is the weight vector, b is the constant term, $\phi(x)$ is the kernel function of transformed data, C is the penalty parameter, and ϵ is error tolerance.

In SVR, the kernel type, C, and ϵ parameters determine the performance of the model and are usually optimized by grid search and cross-validation.

Random Forest Regression

Random Forest is an ensemble-based regression method that is constructed by combining a large number of decision trees. Each tree is trained with a random subset of the dataset and randomly selected variables. The final prediction is obtained by averaging the predictions of all trees [19].

The basic mathematical structure of Random Forest regression:

$$\hat{y} = \frac{1}{N_{trees}}\sum_{i=1}^{N_{trees}} T_i(x)$$

Where, N_{trees} is the total number of trees, $T_i(x)$ is the prediction of the *i*th decision tree. The main parameters optimized in Random Forest are hyperparameters such as the number of trees, maximum depth, and maximum number of variables to be used at each node. These parameters are usually determined by grid search and cross-validation.

D. Evaluation of Model Performance

Objective assessment of the predictive performance of regression models is critical to determine the reliability and practical applicability of the models developed. For this purpose, various statistical metrics are used to measure model performance. In this study, Root Mean Square Error (RMSE) and Coefficient of Determination (R^2) which are the most common and meaningful performance indicators, are used. Each metric assesses a different aspect of the model and is presented below with their mathematical definitions and interpretations.

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

Where y_i is the true (reference) value, $\hat{y}_i$ is the value predicted by the model, *n* is the total number of samples. A low RMSE value indicates that the model predicts with high accuracy. However, since RMSE is the absolute error, it is not appropriate to make direct comparisons between different data sets or variables.

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

Here, $\bar{y}$ is the average of the true values. An R^2 close to 1 indicates that the model fits the data very well; close to 0 indicates that the model is inadequate. Negative R^2 values mean that the model predicts even worse than the average.

III. RESULTS AND DISCUSSION

In this section, the results obtained by combinations of different pre-processing methods and regression models for the estimation of milk fat content are presented. Model performances are mainly evaluated based on R^2 and RMSE values. The findings reveal the effects of pre-processing methods and regression models on prediction accuracy. The obtained results are given in Table 1.

In the modelling using raw spectral data, the PLSR model (R^2 = 0.969) achieved the highest R^2 value. This was followed by Ridge (R^2 = 0.955), Random Forest (R^2 = 0.849), SVR (R^2 = 0.815), and Lasso (R^2 = 0.801) models. When the RMSE values are analyzed, it is seen that the PLSR model shows the best performance with the lowest error (RMSE = 0.106). These results once again confirm the superior performance of PLSR on multicollinearity and high dimensional spectral data. The Lasso and SVR models, on the other hand, show lower prediction performance compared to the other models with low R^2 and high RMSE values in the raw data.

When SNV pre-processing was applied, a significant performance improvement was observed for all models. The PLSR and Ridge models achieved the highest coefficient of determination with R^2 = 0.987 and R^2 = 0.986, respectively. The RMSE values of these models are also quite low (0.068 for PLSR and 0.069 for Ridge). SNV effectively eliminated scattering effects and baseline shifts in the spectral data, allowing the model to capture variations better based on chemical information. SVR and Random Forest

models also showed an increase in R^2 values and a decrease in RMSE after SNV. The Lasso model, however, continued to underperform compared to the other models.

Table 1. The obtained results

Pre-processing Method	Calibration Model	R^2	RMSE		Pre-processing Method	Calibration Model	R^2	RMSE
Raw	PLSR	0.969225	0.106048		SG Smooth	PLSR	0.97416	0.097174
	Ridge	0.955139	0.128038			Ridge	0.956414	0.126205
	Lasso	0.801446	0.269368			Lasso	0.823046	0.254294
	SVR	0.815456	0.25969			SVR	0.821497	0.255405
	Random Forest	0.849092	0.234835			Random Forest	0.851871	0.232662
SNV	PLSR	0.987425	0.06779		SG_Deriv-1	PLSR	0.984722	0.074721
	Ridge	0.98689	0.069216			Ridge	0.986499	0.070241
	Lasso	0.829537	0.249586			Lasso	0.979676	0.086182
	SVR	0.915379	0.175852			SVR	0.981517	0.082185
	Random Forest	0.868198	0.219466			Random Forest	0.958496	0.123154
MSC	PLSR	0.910791	0.180555		SG_Deriv-2	PLSR	0.986165	0.071105
	Ridge	0.957921	0.124005			Ridge	0.986856	0.069305
	Lasso	0.827875	0.2508			Lasso	0.985329	0.07322
	SVR	0.818865	0.257281			SVR	0.982912	0.079023
	Random Forest	0.815254	0.259833			Random Forest	0.977537	0.090603

When MSC pre-processing was applied, the Ridge model (R^2 = 0.958, RMSE = 0.124) and the PLSR model (R^2 = 0.911, RMSE = 0.181) stood out. MSC was particularly effective in reducing scattering variations due to physical matrix differences, but a lower R^2 value was obtained in the PLSR model compared to SNV. In the Lasso, SVR, and Random Forest models, the performance improvement after MSC was limited. This shows that MSC is not able to eliminate spectral variations sufficiently in some models.

When SG smoothing (window size:15, order:2) was applied, the PLSR model (R^2 = 0.974, RMSE = 0.097) and the Ridge model (R^2 = 0.956, RMSE = 0.126) again gave the best results. SG_smoothing reduced the high-frequency noise in the spectral data, allowing the model to capture the main chemical variations better. For the Lasso, SVR, and Random Forest models, the performance improvement was limited.

When the first derivative with SG smoothing (window size:15, order:2) pre-processing method was applied, Ridge (R^2 = 0.986, RMSE = 0.070) and PLSR (R^2 = 0.985, RMSE = 0.075) models stood out. The first derivative improved the performance of linear models, in particular by eliminating baseline shifts and fixed offsets. A significant improvement was also observed in the Lasso and SVR models.

The second derivative with SG smoothing (window size:15, order:2) pre-processing method provided high performance in PLSR (R^2 = 0.983, RMSE = 0.079) and Ridge (R^2 = 0.984, RMSE = 0.076) models. However, the application of the second derivative led to performance degradation in some models, especially Lasso and Random Forest, as it may increase the noise in the spectral data.

The findings show that combinations of pre-processing methods and regression models play a decisive role in model success. In particular, when SNV and first derivative pre-processing methods were used in combination with PLSR and Ridge models, the highest accuracy and lowest error values were achieved in the estimation of milk fat content. These results are in line with the studies in the literature that show that linear models such as PLSR and Ridge models show high performance when used with appropriate pre-processing methods in the estimation of milk components by NIR spectroscopy [20], [21].

Nonlinear models such as SVR and Random Forest, on the other hand, were less affected by pre-processing methods and generally showed lower performance compared to linear models. The Lasso model, despite the advantage of variable selection, provided lower accuracy than other models on high-dimensional spectral data.

This study shows that milk fat can be estimated quickly and reliably by NIR spectroscopy. In particular, the combination of SNV or first derivative pre-processing with PLSR or Ridge models offers high accuracy and reliability in industrial applications. This approach has significant potential for automation and real-time monitoring of quality control processes in the dairy industry.

IV. CONCLUSION

This study aimed to estimate the fat content of milk quickly and reliably using NIR spectroscopy and various chemometric methods. The results obtained with different combinations of pre-processing methods (SNV, MSC, SG, First and Second Derivative) and regression models (PLSR, Ridge, SVR, Lasso, Random Forest) showed that the pre-processing methods significantly affected the model performance. In particular, the use of SNV and first derivative pre-processing methods in combination with PLSR and Ridge models resulted in the highest accuracy and lowest error values. The findings support NIR spectroscopy as a promising tool for automation and real-time monitoring of quality control processes in the dairy industry. Future research could focus on improving the generalizability and robustness of the model by applying different chemometric approaches and model validation techniques on larger and more diverse milk samples.

ACKNOWLEDGMENT

This study was not supported by any organization.

REFERENCES

[1] A. Haug, A. T. Høstmark, and O. M. Harstad, “Bovine milk in human nutrition - A review,” *Lipids Health Dis*, vol. 6, 2007, doi: 10.1186/1476-511X-6-25,.

[2] R. G. Jensen, “The Composition of Bovine Milk Lipids: January 1995 to December 2000,” *J. Dairy Sci*, vol. 85, pp. 295–350, 2002, doi: 10.3168/jds.S0022-0302(02)74079-4.

[3] D. M. Barbano, Y. Ma, and M. V. Santos, “Influence of raw milk quality on fluid milk shelf life.,” *J Dairy Sci*, vol. 89 Suppl 1, pp. E15–E19, Mar. 2006, doi: 10.3168/jds.s0022-0302(06)72360-8.

[4] B. T. Kao, K. A. Lewis, E. J. DePeters, and A. L. Van Eenennaam, “Endogenous production and elevated levels of long-chain n-3 fatty acids in the milk of transgenic mice,” *J Dairy Sci*, vol. 89, no. 8, pp. 3195–3201, Aug. 2006, doi: 10.3168/jds.S0022-0302(06)72594-2.

[5] R. Karoui and J. De Baerdemaeker, “A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products,” *Food Chem*, vol. 102, no. 3, pp. 621–640, Jan. 2007, doi: 10.1016/J.FOODCHEM.2006.05.042.

[6] B. M. Nicolaï *et al.*, “Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review,” *Postharvest Biol Technol*, vol. 46, no. 2, pp. 99–118, Nov. 2007, doi: 10.1016/J.POSTHARVBIO.2007.06.024.

[7] R. Tsenkova, S. Atanassova, K. Toyoda, Y. Ozaki, K. Itoh, and T. Fearn, “Near-Infrared Spectroscopy for Dairy Management: Measurement of Unhomogenized Milk Composition,” *J Dairy Sci*, vol. 82, no. 11, pp. 2344–2351, Nov. 1999, doi: 10.3168/JDS.S0022-0302(99)75484-6.

[8] L. R. Joppa, G. A. Hareland, and R. G. Cantrell, “Quality Characteristics of the Langdon Durum-dicoccoides Chromosome Substitution Lines,” *Crop Sci*, vol. 31, no. 6, pp. 1513–1517, Nov. 1991, doi: 10.2135/CROPSCI1991.0011183X003100060024X.

[9] J. A. Diaz-Olivares *et al.*, “Near-infrared spectra dataset of milk composition in transmittance mode,” *Data Brief*, vol. 51, p. 109767, Dec. 2023, doi: 10.1016/J.DIB.2023.109767.

[10] G. K. Krug, “COMPUTER AIDED DESIGN OF EXPERIMENTS.,” *Periodica Polytechnica Electrical Engineering*, vol. 19, no. 3, pp. 181–189, 1975, doi: 10.2307/1266770.

[11] Å. Rinnan, F. van den Berg, and S. B. Engelsen, “Review of the most common pre-processing techniques for near-infrared spectra,” *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, Nov. 2009, doi: 10.1016/J.TRAC.2009.07.007.

[12] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Appl Spectrosc*, vol. 43, no. 5, pp. 772–777, 1989, doi: 10.1366/0003702894202201;JOURNAL:JOURNAL:ASPC;CTYPE:STRING:JOURNAL.

[13] P. Geladi, D. MacDougall, and H. Martens, "Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat," *Appl Spectrosc*, vol. 39, no. 3, pp. 491–500, 1985, doi: 10.1366/0003702854248656.

[14] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Anal Chem*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964, doi: 10.1021/AC60214A047/ASSET/AC60214A047.FP.PNG_V03.

[15] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, Oct. 2001, doi: 10.1016/S0169-7439(01)00155-1.

[16] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," vol. 12, no. 1.

[17] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[18] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support Vector Regression Machines," *Adv Neural Inf Process Syst*, vol. 9, 1996.

[19] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[20] J. A. Diaz-Olivares, I. Adriaens, E. Stevens, W. Saeys, and B. Aernouts, "Online milk composition analysis with an on-farm near-infrared sensor," *Comput Electron Agric*, vol. 178, p. 105734, Nov. 2020, doi: 10.1016/J.COMPAG.2020.105734.

[21] V. Fonseca Diaz, B. De Ketelaere, B. Aernouts, and W. Saeys, "Cost-efficient unsupervised sample selection for multivariate calibration," *Chemometrics and Intelligent Laboratory Systems*, vol. 215, p. 104352, Aug. 2021, doi: 10.1016/J.CHEMOLAB.2021.104352.