foods

MDPI

Article

Establishment and Validation of Fourier Transform Infrared Spectroscopy (FT-MIR) Methodology for the Detection of Linoleic Acid in Buffalo Milk

Zhiqiu Yao [1,2], Pei Nie [1,3], Xinxin Zhang [1,2], Chao Chen [1,2], Zhigao An [1,2], Ke Wei [1,2], Junwei Zhao [1,2], Haimiao Lv [1,2], Kaifeng Niu [1,2], Ying Yang [1,2], Wenna Zou [1,2] and Liguo Yang [1,2,*]

1 National Center for International Research on Animal Genetics, Breeding and Reproduction (NCIRAGBR), Ministry of Science and Technology of the People's Republic of China, Huazhong Agricultural University, Wuhan 430070, China
2 Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Education, College of Animal Science and Technology, Huazhong Agricultural University, Wuhan 430070, China
3 College of Veterinary Medicine, Hunan Agricultural University, Changsha 410128, China
* Correspondence: ylg@mail.hzau.edu.cn

Abstract: Buffalo milk is a dairy product that is considered to have a higher nutritional value compared to cow's milk. Linoleic acid (LA) is an essential fatty acid that is important for human health. This study aimed to investigate and validate the use of Fourier transform mid-infrared spectroscopy (FT-MIR) for the quantification of the linoleic acid in buffalo milk. Three machine learning models were used to predict linoleic acid content, and random forest was employed to select the most important subset of spectra for improved model performance. The validity of the FT-MIR methods was evaluated in accordance with ICH Q2 (R1) guidelines using the accuracy profile method, and the precision, the accuracy, and the limit of quantification were determined. The results showed that Fourier transform infrared spectroscopy is a suitable technique for the analysis of linoleic acid, with a lower limit of quantification of 0.15 mg/mL milk. Our results showed that FT-MIR spectroscopy is a viable method for LA concentration analysis.

Keywords: FT-MIR; linoleic acid; machine learning; accuracy profile

Citation: Yao, Z.; Nie, P.; Zhang, X.; Chen, C.; An, Z.; Wei, K.; Zhao, J.; Lv, H.; Niu, K.; Yang, Y.; et al. Establishment and Validation of Fourier Transform Infrared Spectroscopy (FT-MIR) Methodology for the Detection of Linoleic Acid in Buffalo Milk. *Foods* **2023**, *12*, 1199. https://doi.org/10.3390/foods12061199

Academic Editor: Ksenija Radotić

Received: 9 February 2023
Revised: 28 February 2023
Accepted: 10 March 2023
Published: 12 March 2023

1. Introduction

Milk is an important route in humans for nutrient intake, and the fatty acids in milk are associated with many biological functions in humans [1]. The dietary intake of fatty acids has an important influence on coronary disease; specifically, saturated fatty acids (SFA) increase serum cholesterol levels, whereas polyunsaturated fatty acids (PUFAs) reduce the risk of coronary disease [2]. In addition, studies have shown that the fatty acids in milk are also related to the technological properties of milk and the processing of dairy products [3]. The composition of milk fat and fatty acid content reflects to a certain extent the health status of the cow [4].

Linoleic acid (LA) is a type of PUFA that has been shown to have various health benefits, including reducing the risk of chronic diseases and improving insulin sensitivity [5,6]. Milk is an important source of LA, which is considered a potential anticarcinogen and can be manipulated through dietary management [7]. As the economy continues to develop, there is a growing demand for milk that is nutritionally valuable. The dairy industry, therefore, faces two major challenges: (1) aligning the fatty acid composition of milk with consumer preferences, and (2) finding reliable and precise methods to quantify the FA composition of milk [8]. The traditional methods for determining LA content in milk products are gas chromatography (GC) [9] or gas chromatography–mass spectrometry [10], which are time-consuming and labor-intensive and often involve the use of harmful chemicals.

Fourier transform mid-infrared spectrometry (FT-MIR) is a widely utilized analytical technique that has been demonstrated to be effective in a range of applications within the dairy industry. Specifically, FT-MIR has been demonstrated to be valuable in the analysis of antibiotics present in milk [11,12], the quantification of fat and protein content in milk [13], the prediction of methane emissions [14], and the prevention of early lactation diseases in cattle [15,16]. FT-MIR has been increasingly used for the analysis of fatty acids in milk due to its advantages of high throughput in real-time, sensitivity, and low sample preparation requirements [17].

In recent years, there has been a growing interest in using FT-MIR in combination with multivariate analysis techniques, such as partial least squares regression (PLSR), to quantify the PUFA content in milk products. Mid-infrared spectroscopy has been widely used in the rapid prediction of fatty acids. PLSR is probably the most widely used technique in spectral analysis. Many researchers have successfully measured the fat, protein, solid non-fat, and fatty acid content in milk using PLSR regression [18–21]. With the development of computational power and machine learning methods, more and more multivariate models are used to calibrate the concentration of components in milk. The principal component regression (PCR) algorithm downscales the original features using principal components analysis (PCA) and performs linear regression on the reduced predictor variables, which are the principal components, to predict the target variable. By utilizing a smaller number of principal components that explain the majority of the variance in the data with respect to the target variable, PCR is more effective in mitigating overfitting than linear regression on all original features, particularly for high-dimensional data such as spectra [22]. In recent studies, artificial neural networks (ANNs) have recently been investigated in FT-MIR spectroscopic analysis [23,24]. Random forests (RF) employ an evaluation of the relevance of variables to selectively choose informative variables, thereby facilitating the construction of models that are both parsimonious and robust, and ultimately enhancing the predictive power [25,26].

Some recent endeavors employing FT-MIR spectroscopy have explored quantifying linoleic acid. Beriain et al. [27] predicted the α-linolenic acid and LA in intramuscular fat by using the ANN algorithm and achieved good forecasted results. In the field of dairy analysis, Bonfatti et al. [28] successfully developed a milk fatty acid prediction model for Italian Simmental cattle using MIR with PLSR algorithm on 1040 milk samples. Similarly, Coppa et al. [21] used GC combined with FT-MIR to develop a fatty acid prediction model for 250 Holstein milk samples. However, although GC is a useful technique for analyzing monounsaturated fatty acids, its accuracy may be reduced when analyzing complex mixtures of polyunsaturated fatty acid methyl esters containing trans double bonds, such as LA and alpha-linolenic acid [29]. In addition, the complexity and cost associated with GC analysis for large numbers of samples and the need for expert operators are important factors to consider. Notably, there have been no previous investigations on the prediction of linoleic acid content in buffalo milk using FT-MIR, and only a limited number of studies have assessed the accuracy and precision of FT-MIR-based predictions of PUFA [30].

Direct or spectral interference is a common issue in chemical analysis based on spectroscopic methods, where the sensor is not perfectly specific for the analyte [31]. Unintended interference can occur, especially when utilizing PLSR for the compositional analysis of highly complex samples [32]. It is important to carefully evaluate and control potential sources of interference in spectroscopic methods to ensure accurate and reliable chemical analysis. Therefore, the objectives of this study were twofold: (1) to modify the FT-MIR method by incorporating the standard addition technique and establish a more streamlined machine learning prediction model for linoleic acid in milk, and (2) to employ a novel validation strategy to evaluate the accuracy and precision of FT-MIR for the determination of linoleic acid in milk.

2. Materials and Methods

2.1. Sampling

Over a 3-month period from April to June 2022, milk samples were collected from 31 buffaloes in Hubei, China. For each sampling day, 50 mL of milk was collected in the morning and another 50 mL in the afternoon, and then mixed into a single sample to reflect changes in milk composition throughout the day. In total, 12 L milk samples were collected and stored at −20 °C for further analysis.

2.2. FT-MIR and Preprocessing Method

To process the samples, they were first rapidly thawed in a 40 °C water bath and then centrifuged at 2 °C, using a refrigerated centrifuge, at 3000 rpm for 15 min to eliminate fat [33]. The composition of skimmed milk was analyzed using the MilkoScan FT-6000 (FOSS Analytical A/S, Hillerød, Denmark), which revealed that the fat content of whey was less than 0.05%.

Linoleic acid (LA) was randomly added to skimmed milk samples in seven different concentrations (1, 5, 10, 20, 50, 70, 100 mg/100 mL milk). Isopropanol was utilized as the diluent for the LA [34]. There were 15 samples for each concentration, and a total of 105 samples were used for FT-MIR analysis. MIR spectra were obtained for each sample using the Milkoscan FT 6000. The acquisition was performed twice, and the results were subsequently averaged. The MIR spectra were recorded in the region between 926 and 5012 cm^{-1} and omitted the O–H bending region (1600–1710 cm^{-1}) and the O–H stretching region (3020–5012 cm^{-1}) due to the high water content in milk [35]. The remaining region (926 to 1618 cm^{-1} and 1705 to 3025 cm^{-1}; 524 data points) was selected for analysis [36].

To further process the raw spectra, 7 different preprocessing methods were applied, including standard normal variate (SNV), 11-point Savitzky–Golay algorithm (SG), first derivative + Savitzky–Golay algorithm (SG-1), second derivative + Savitzky–Golay algorithm (SG-2), SNV + Savitzky–Golay algorithm (S-SG), SNV + Savitzky–Golay algorithm + first derivative (S-SG-1), and SNV + Savitzky–Golay algorithm + second derivative (S-SG-2) (Figure 1). The R packages "prospect" (version 0.26) and "baseline" (version 1.3-4) were utilized for the preprocessing steps.

2.3. Machine Learning Algorithms

In this study, we aimed to determine the optimal quantitative model for the estimation of linoleic acid in milk using various machine learning techniques. All the machine learning algorithms utilized the CARET package version 6.0–93 in R program (version 4.2.2 https://www.r-project.org/ (accessed on 8 September 2022)) [35].

The FT-MIR data (n = 105) was randomly divided into a training set (80%) and a test set (20%) for building and validating the models, respectively. The numerical parameters for each model were entered using the "expand.grid" function and optimized using cross-validation (CV) statistics. We selected the model with the lowest root mean square error of cross-validation ($RMSE_{CV}$) from all preprocessing methods. PLSR is a widely utilized chemometric method in the analysis of spectroscopic data, utilizing latent variables (LV) to decompose the spectral data into systematic variations that account for the observed variance [37]. In comparison, the latent variable of PCR is the number of principal components and the minimum number of principal components required to explain 95% of the variance [38]. ANNs represent a nonlinear extension of traditional linear regression models [39]. While linear regression is limited to modeling linear relationships between features and targets, ANNs have the capability to model complex nonlinear relationships through the utilization of hidden layers [40]. Regularization techniques play a crucial role in preventing overfitting in ANN models, thus improving their accuracy on novel data sets [41]. In the context of the CARET package, the parameter "size" refers to the number of units in a hidden layer, and the parameter "decay" represents the regularization strength. For PLSR and PCR, the maximum number of latent variables was set to 25. The number of the hidden layer for the ANN was varied from 1 to 5, and the decay values were tested for 0,

0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, and 0.5. The performance of each model was evaluated using internal 10-fold cross-validation statistics, including $RMSE_{CV}$ and coefficients of determination (R2 cv). The models were then validated by estimating RMSE of prediction (RMSEP) on the external test set.

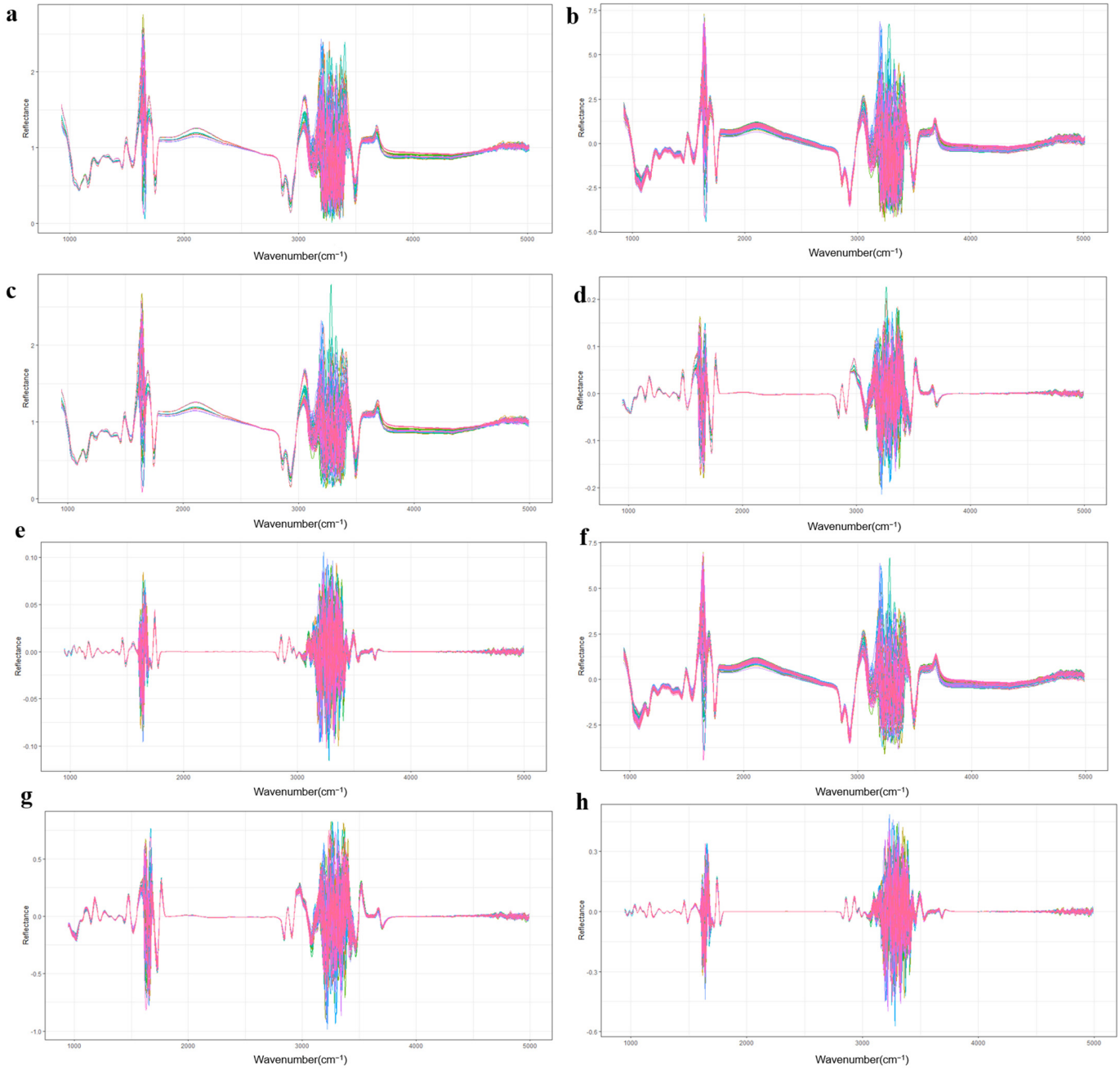

Figure 1. Mid-infrared spectra after various preprocessing methods. (**a**) Raw spectra; (**b**) Mid-infrared spectra after standard normal variables processing; (**c**) Mid-infrared spectra after Savitzky–Golay algorithm processing; (**d**) Mid-infrared spectra after first derivative and Savitzky–Golay algorithm processing; (**e**) Mid-infrared spectra after second derivative and Savitzky–Golay algorithm processing; (**f**) Mid-infrared spectra after SNV and Savitzky–Golay algorithm processing; (**g**) Mid-infrared spectra after SNV, Savitzky–Golay algorithm, and first derivative processing; (**h**) Mid-infrared spectra after SNV, second derivative, and Savitzky–Golay algorithm processing.

Random forests (RF) have been demonstrated to hold promise in the realm of feature selection [42]. This procedure involves creating a random forest model and then

performing 1000 iterations. Through the creation of a random forest model and subsequent iterations, the importance scores of features were evaluated based on the accuracy of model predictions of the target variable (LA) after replacing the response variable (spectral bands). Spectral bands that are more predictive of the outcome will have relatively high importance scores in each run, while other spectral bands with lower predictivity will only have randomly importance scores. This process enables the significance of features to be calculated [43]. In this study, we employed the rfPermute package (version 2.5.1) in R to perform variable selection using RF. The number of trees utilized in the RF model was set at 500 [44]. The PLS, PCR, and ANN models were again constructed by selecting spectral regions with significance levels less than 0.05. These models underwent variable optimization and performance evaluation in a manner consistent with the methodology previously described.

2.4. Quality Control for the Method

The developed method underwent validation in accordance with the International Conference on Harmonization (ICH) Q2 (R1) guidelines. The Limit of Detection (LOD) was determined by utilizing 10 skimmed milk samples, and calculating the standard deviation of the matrices. The LOD was determined as three times the standard deviation of the ten sample [45].

$$\mathrm{LOD} = 3 \times S_0 \tag{1}$$

S_0 is the estimated standard deviation of single results at zero concentration.

In terms of relative bias, recovery, repeatability, intermediate precision, lower limit of quantification (LLOQ) and upper limit of quantification (ULOQ), the validation protocol employed a $3 \times 5 \times 3$ ($i \times k \times j$) full factorial experiment design [46]. Five different concentration levels (k) of linoleic acid (5 mg/100 mL, 10 mg/100 mL, 20 mg/100 mL, 50 mg/100 mL and 100 mg/100 mL) were investigated, with each level being conducted in three replicates (i) on three different days (j), resulting in a total of 45 samples [46].

The trueness of the method was evaluated through the expression of Bias and Recovery.

$$\mathrm{Bias}(\%) = \frac{\bar{Y} - Y_r}{Y_r} \times 100 \tag{2}$$

$$\mathrm{Recovery}(\%) = \frac{\bar{Y}}{Y_r} \times 100 \tag{3}$$

where Y_r is the theoretical value, $\bar{Y}$ is the average value of a series of measurements.

Precision is evaluated at two levels: repeatability and intermediate precision. This requires the calculation of the mean square of inter-series (MSB) and intra-series (MSE) [47].

If MSE < MSB, then:

$$\text{Repeatability} : {\sigma_{Re}}^2 = \mathrm{MSE} \tag{4}$$

$$\text{Intermediate precision} : {\sigma_{In}}^2 = \frac{\mathrm{MSB} - \mathrm{MSE}}{n} \tag{5}$$

Otherwise:

$$\text{Intermediate precision} = \text{Repeatability} = \frac{1}{mn-1} \sum_{i=1}^{m} \sum_{j=1}^{n} (Y_{ij} - \bar{Y})^2 \tag{6}$$

where Y_{ij} is the average of the calculated concentration of the j-th concentration level of the i-th series; m is the number of days; n is the number of replicates per series.

The current assay acceptance criteria in practice require that at least four out of six samples have an observed mean at the lower limit of quantitation (LLOQ) within 20% of the theoretical value ($\beta = 4/6 \approx 66.7\%$), and the observed precision to be $\leq$20% coefficient of variation [48].

The accuracy profile based on β-content tolerance intervals is a powerful tool for method validation and quality control [49]. This ideal acceptance criterion would ensure that a high proportion (β = 66.7%) of future observations lie within acceptance limits (±20%), with a high level of confidence (confidence level = 0.9). By calculating the lower limit (*L*) and upper limit (*U*) of the tolerance at a particular concentration, the tolerance of the measured value of a specified proportion (β) of all samples will be within the interval [*L*,*U*] with the specified confidence level. It can be considered that if the number of observed values Y_{n+1} within the tolerance interval [*L*,*U*] accounts for more than 0.667 of the total Y_n, and the confidence level is 0.9, then the detection method is valid, and the formula is shown as follows.

$$\text{Confidence level} = P[P[L \leq Y_{n+1} \geq U/Y_n] \geq \beta] \tag{7}$$

For instance, for β = 0.667 and confidence level is 0.9, the β- content tolerance interval represents a 90% probability (*p*) that 66.7% of the individual observations of the population are included in the interval [*L*,*U*] [50]. The determination is accepted if the resulting tolerance limits *L* and *U* are completely within acceptance limits (±20%) of the theoretical value; Otherwise, it's not.

According to Kulkarni's approach [51], the tolerance interval [*L*,*U*] can be rewritten into the following form:

$$[L(\%), U(\%)] = [\text{bias}(\%) - \chi_k \times \text{RSD}(\%), \text{bias}(\%) + \chi_k \times \text{RSD}(\%)] \tag{8}$$

Where:

$$\text{RSD}(\%) = \frac{\sigma_{In}}{Y_r} \times 100 \tag{9}$$

$$\chi_k = \sqrt[2]{\frac{k \times \chi^2_{1;0.667}(\lambda)}{\chi^2_{k;0.9}(0.1)}} \tag{10}$$

$\chi^2_{1;0.667}\tau$ is the 66.7th quantile of a noncentral chi-square distribution with the degree of freedom 1. λ is the noncentrality parameter. $\chi^2_{f';0.9}(0.1)$ is the 90th quantile of a noncentral chi-square distribution with the degree of freedom *k*. χ_k denotes the chi-square distribution associated with the variable *k* [50,52].

$$k = \frac{(R'+1)^2}{\left(R' + \frac{1}{n}\right)^2/(m-1) + \left(1 - \frac{1}{n}\right)/mn} \tag{11}$$

$$\lambda = \frac{nR'+1}{mn(R'+1)} \tag{12}$$

$$R' = MAX\left[0, \frac{1}{n}\left(\frac{\text{MSB/MSE}}{F_{0.85}(m(n-1)); (m-1)} - 1\right)\right] \tag{13}$$

$F_{0.85}(m(n-1)); (m-1)$ is the 85th percentile value of F distribution with the degree of freedom $m(n-1)$ and $m-1$. The concentration at which the tolerance interval is less than acceptance limits is the limit of quantification. The acceptance limits is typically set at ±20% [52,53].

The procedure for building an accuracy file can be outlined as follows:

(1) Calculate the β-content tolerance interval at a confidence level of 0.9 for each concentration level using Equations (13) or (8), resulting in a lower and upper limit for the interval, denoted as [*L*,*U*].
(2) Graphically represent the results in a 2D plot, with the concentration level plotted on the horizontal axis and the tolerance interval limits (*L*,*U*) plotted on the vertical axis.
(3) Compare the tolerance interval limits (*L*,*U*) with the acceptance limits of −20% to +20% around the theoretical value. If the tolerance interval falls entirely within

this acceptance range, the analytical method is deemed valid for the corresponding concentration level. However, if the tolerance interval exceeds these limits, the method is not accepted for use at that concentration level.

3. Results and Discussion

3.1. Set Up of the Prediction Models

First, we implemented a 10-fold cross-validation process to avoid overfitting. Cross-validation has proven to be a good method for model resampling and is widely used for the mid-infrared prediction of milk composition [54,55]. The performances of the various methods (PLSR, PCR, and ANN) are summarized in Table 1, with the RMSE and coefficient of determination (R^2). The best model was determined by the smallest RMSE and highest R^2. The $RMSE_{CV}$ values for the PLSR, PCR, and ANN models were all found to be below ten. In our study, the $RMSE_{CV}$ values of the training set were always lower than the one observed for test set, as mentioned by Soyeurt and Grelet [35]. Results showed that $RMSE_{CV}$ values for PLSR, PCR, and ANN were similar, ranging from 5.1 mg/100 mL–7.3 mg/100 mL, with R^2_{CV} values also globally similar and ranging from 0.96–0.98. This indicates that the predictive performance of the three models is similar. There were also some differences in correlation values between predictions on the test set. Higher correlation was observed between the predictions given by PLSR and PCR(0.98) compared to those given by ANNs. Our analysis revealed that the PCR method outperformed the PLSR method, with slightly higher predictive accuracy. This difference in performance may be attributed to the distinct component extraction processes employed by PCR and PLSR. Specifically, PLSR identifies regressors from predictors that maximize the covariance with the response variable, while PCR employs principal component analysis (PCA) to identify the direction of greatest variability in the predictor variables and project them into a low-dimensional space to form principal components, which are subsequently used to explain the response variable. The component extraction step in PCR is capable of identifying superior candidate regression components by meticulously scrutinizing the covariance structure among the predictor variables, which may be overlooked by PLSR. Such phenomena have been observed in previous studies as well [56,57]. It is worth noting that the performance of different methods depends on the nature of the analyzed data and the data processing methods used. This is one of the reasons why it is not recommended to use the same milk fatty acid prediction model across different species.

Table 1. Performance of 10-fold cross-validation and external validation for predicting LA in milk using 5 different machine learning algorithms [1].

	Pre-Processing	LV	$RMSE_{CV}$	$RMSE_{CV}$ SD	R^2_{CV}	RMSEp	R^2_p
PLSR	SG-1	nLV [2] = 20	7.325	0.546	0.958	4.094	0.984
PCR	SG	Nlv = 17	5.198	0.725	0.980	3.662	0.987
ANN	SNV-SG-1	Size [3] = 5 Decay = 0.3	6.274	1.809	0.963	6.426	0.961

[1]. PLSR = partial least squares regression; PCR = principle component regression; ANN = artificial neural network; $RMSE_{CV}$ = root mean square error in cross-validation; R^2cv = cross-validation R^2; $RMSE_{CV}$ SD = standard deviation of $RMSE_{CV}$; R^2_P = R2 in prediction; $RMSE_P$ = RMSE in prediction. [2]. nLV = number of latent variables. [3]. Size = the number of nodes in the hidden layer; decay = the penalty used for ANN.

3.2. Models Built with the Spectral Regions Selected by RF

Our research on importance measures in random forests has focused on finding data points where the predictor variables are highly correlated. The application of RF results in a significant reduction in the number of variables in each model. The number of data points dropped from 524 before the selection to 135.

RF is widely used to assess the importance of features. Wang et al. [58] employed RF feature selection to investigate the relative contributions of soil factors, microbial parameters, and climatic factors in altering soil organic carbon levels. Chen et al. [59] used RF to

evaluate the most significant drivers of soil fungal diversity, including plant communities and soil physicochemical properties. Similarly, Andreas et al. [60] utilized RF to identify tillage type as the most important factor affecting dairy cows with Fasciola hepatica, with higher-ranking variables yielding more accurate predictions than those with lower importance scores. The variable importance measure can be used by RF to select and order the spectral regions that are most predictive. Usually, MIR data points are ranked according to decreasing asymptotic *p*-values and importance value. The process of changing random number seeds will result in slightly different results for random forests [61]. Therefore, the response variable was permuted 1000 times to generate new RF models, and the data points that were most correlated with linoleic acid and significant at a *p*-value of less than 0.05 were selected for modeling [62,63]. Most of the selected data points were included in the spectral subsets 940–1215 cm^{-1}, 1342–1489 cm^{-1}, 2364–2399 cm^{-1}, 2823–2935 cm^{-1}, and 3715–3846 cm^{-1} (Figure 2). These regions are highly correlated with fatty acids. The first region (940–1215 cm^{-1}) is related to the asymmetric vibrations of the C-O-C group in esters; the second region (1342–1489 cm^{-1}) is characteristic of the C = O ester Fermi resonance; the third region (1720–1766 cm^{-1}) is characteristic of the stretching vibrations of the carbonyl group in esters; the fourth region (2350–2357 cm^{-1}) is a synergistic region associated with fatty acids and has been shown to assist in the prediction of fatty acids in milk to some extent [30]. The fourth region (2823–2935 cm^{-1}) is characteristic of C-H stretching absorption [64].

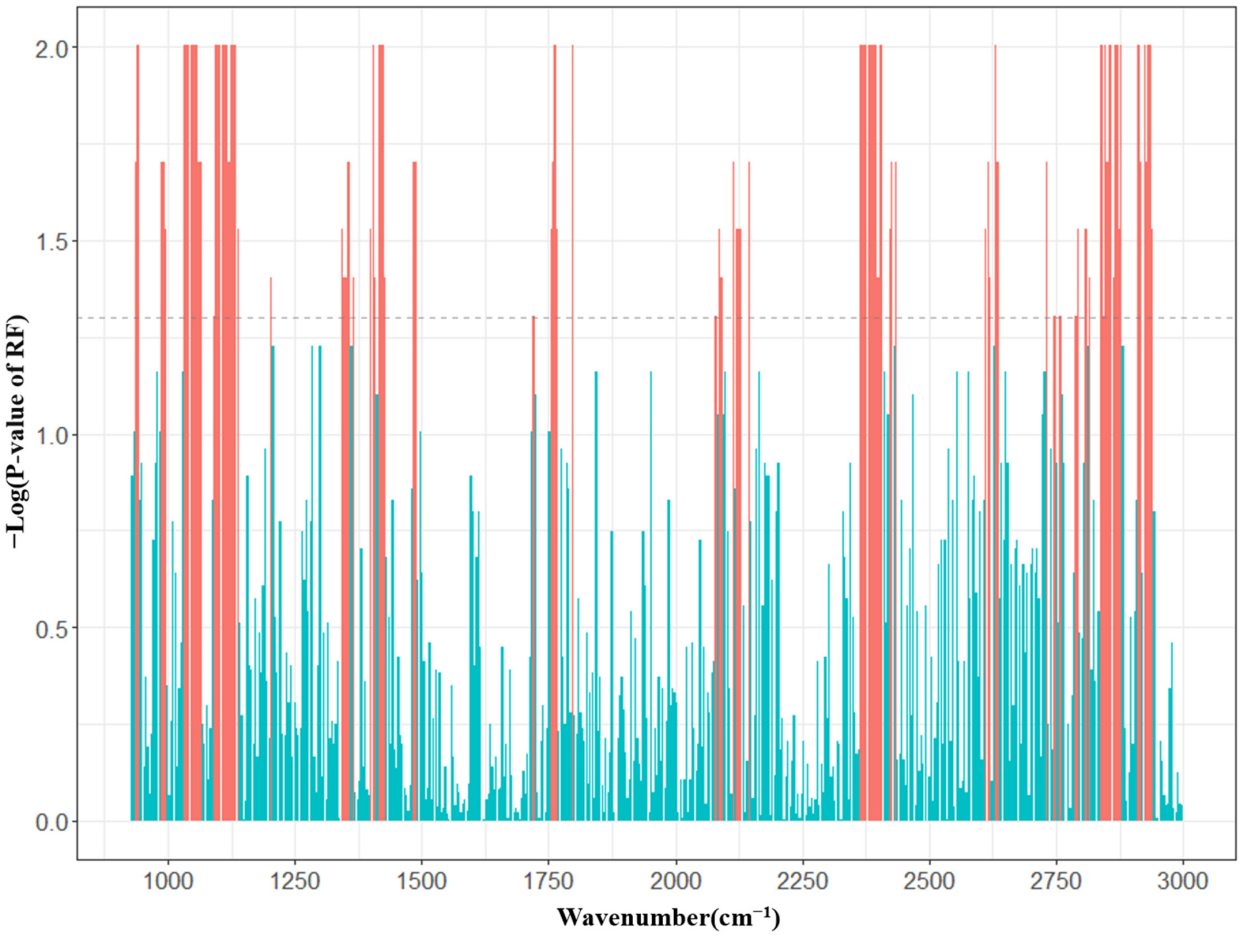

Figure 2. Spectral regions selected by RF for the prediction of linoleic acid. The red domains indicate a significant association with LA ($p \leq 0.05$), and the green domains indicate a non-significant association with LA ($p > 0.05$).

The validation results for the prediction models built using the wavelengths selected by RF are presented in Table 2, along with the number of latent variables. The results indicate that compared to the full spectrum model, the $RMSE_{CV}$ value of the RF model is generally lower. The application of RF resulted in a reduction of the R^2_P value by 0.2% in the ANN mode, while the R^2_P value increased by 0.3% in the PCR model. No differences were observed in the R^2_P value of the PLSR model, but the R^2_{cv} value of PLSR increased by 2.1% after RF feature selection. Thus, the application of RF has produced simpler models, and the predictive power of these simplified models is comparable to that of full spectrum models. As mentioned above, the performance of a method is closely related to the characteristics of the data set, the preprocessing methods, and the relationship between the predictor and response variables. After using RF to extract the original features, the performance of the PLS method was slightly better than that of the PCR method.

Table 2. Performance of 10-fold cross-validation and external validation of LA in predicted milk based on 5 different machine learning algorithms after random forest algorithm variable selection [1].

	Pre-Processing	LV	$RMSE_{CV}$	$RMSE_{CV}$ SD	R^2_{CV}	$RMSE_P$	R^2_P
PLSR	SG	nLV [2] = 8	4.714	1.188	0.983	4.113	0.984
PCR	SG	nLV = 14	5.669	0.836	0.976	4.161	0.983
ANN	SNV	Size [3] = 6 Decay = 0.4	7.616	2.373	0.951	6.566	0.959

[1]. PLSR = partial least squares regression; PCR = principle component regression; ANN = artificial neural network. $RMSE_{CV}$ = root mean square error in cross-validation; R^2cv = cross-validation R^2; $RMSE_{CV}$ SD = standard deviation of $RMSE_{CV}$; R^2_P = R2 in prediction; $RMSE_P$ = RMSE in prediction. [2]. nLV = number of latent variables. [3]. Size = the number of nodes in the hidden layer; decay = the penalty used for ANN.

In detail, lower $RMSE_P$ values were observed between the predictions given by the PLSR and PCR (4.1) compared with ANNs (6.5). This suggests that nonlinear methods, such as ANN models, were not suitable, but linear PLSR showed good performance. Previous research has suggested that FT-MIR predictions with partial least square models are promising approaches [65,66]. This is in agreement with Soyeurt et al. [35], showing that PLSR has better predictive performance than ANNs in orange variety classification. Improving the performance of ANNs requires a large training data set to learn complex data interactions by tuning its hyperparameters, such as size and decay, in this study [67]. In addition, epochs, activation function, and learning rate all affect the predicting capabilities of the ANN. From our results, the ANN does not seem to perform well when the training population is small. The RMSE values for the other four linear models are smaller than those with the ANN, which also suggests that the complex nonlinear relationship between the predictors and target traits is limited [68,69]. It was evident from the values of R^2 and RMSE that, even though all three models (PLSR, PCR, and SVR) fitted well to the experimental design, PLSR offered better predictive and approximation accuracy. The best predictive performance was achieved by the PLSR with the mean R^2_P value of 0.984 and a RMSEp value of 4.113 mg/100 mL.

We observed a higher predictive ability for linoleic acid content compared with previous studies on FT-MIR predictions, which obtained R^2 values ranging from 0.43–0.89 [30]. This improvement could be attributed to the expression of fatty acid content estimated in g/100 mL milk, which is more accurate than g/100 g FA [21,70]. Additionally, the utilization of the standard addition method, which is commonly used in the method validation of other analytical methods such as GC, was instrumental in avoiding interference from other fatty acids [34]. To the best of our knowledge, it was the first time that the RF method was used on FT-MIR data to select salient features. Although this method appears promising, further studies will be needed to fully understand its limitations.

3.3. MIR Method Validation

The method detection limit for LA was determined using FT-MIR spectroscopy according to the ICH Q2 (R1). The LOD was found to be 3.42 mg/100 mL based on the standard deviation of the blank sample signals (n = 10). The FT-MIR method was validated towards recovery, repeatability, intermediate precision, range, and accuracy for the quantification of LA according to the ICH Q2 (R1). The acceptable limits were set at ±20% for the IR method [53,71].

The trueness represents the closeness of the average to the true value, and precision is the closeness among a series of measurements [72]. The uncertainty is a dispersion of measured values from the expected value [73]. The total uncertainty includes the random error and the systematic error.

Table 3 illustrates that the results for LA at 20, 50, and 100 mg/100 mL concentration levels have good relative bias and recovery, with relative bias ranging from −0.56% to 2.15% and recovery ranging from 99.44% to 102.47%. The repeatability and intermediate precision of LA at 10 mg/100 mL and 20 mg/100 mL concentrations were 1.89% and 2.07% respectively. The repeatability and inter-assay precision of LA at 50 mg/100 mL were 8.34% and 9.46% respectively, while the repeatability and inter-assay precision of LA at 100 mg/100 mL were 11.76% and 13.61%, respectively. In our results, the intermediate precision is worse than the repeatability, which means that there is an effect of day-to-day variability on the spectral data at these concentration levels [74].

Table 3. Trueness, precision, and accuracy results for each concentration level in the validation data [1].

		Trueness		Precision		Accuracy	
Level (mg/100 mL)	**Mean Calculated Concentration [2] (mg/100 mL)**	**Relative Bias (%)**	**Recovery (%)**	**Repeatability (%)**	**Intermediate Precision (%)**	**Relative β-Expectation Tolerance Limits (%) [3]**	**β-Expectation Tolerance Limits (mg/100 mL) [4]**
5	6.52 ± 1.72	30.4	130.4	3.89	8.21	[−59.18, 119.99]	[2.04, 10.99]
10	10.67 ± 1.29	6.7	106.7	1.89	1.89	[−14.8, 28.21]	[8.51, 12.82]
20	20.43 ± 1.35	2.15	102.15	2.07	2.07	[−9.11, 13.41]	[18.17, 22.68]
50	49.72 ± 2.85	0.56	99.44	8.34	9.46	[−10.17, 9.07]	[44.91, 54.53]
100	102.02 ± 3.42	2.63	102.63	11.76	13.61	[−10.24, 14.96]	[96.05, 108.66]

[1]. The validation criteria are based on the ICH guide to the validation of analytical methods. [2]. Mean ± SD. [3]. The β-CTI (%) is the relative β-content tolerance interval. [4]. Abs β-CTI is the absolute β-content tolerance interval.

The accuracy of LA at 20 mg/100 mL (−9.11, 13.41), 50 mg/100 mL (−10.17, 9.07), and 100 mg/100 mL (−10.24, 14.96) concentration were found to be within the acceptable range of −20 to 20%. The accuracy of 5 mg and 10 mg/100 mL level were outside the acceptance limits. This suggests that systematic and random errors increase as the concentration level decreases [46].

As shown in Figure 3a, the relationship between the predicted concentrations and the true concentrations was evaluated by the linear equation: $y = 1.013x + 0.4959$ with R^2 of 0.9948. The slope and R^2 values of the linear equation demonstrate the good agreement between the MIR predictions and the theoretical values.

The accuracy profile is a pictorial tool that is widely used for the quality control of medicines [75]. LLOQ and ULOQ are the lowest and highest concentration levels where the β-tolerance expectation limits are included within the acceptable limits. In our study, the LLOQ value was 15.54 mg/100 mL, and the ULOQ value was 100 mg/100 mL.

In this study the acceptable limit was set at ±20%, and in the other literature the acceptable limit has been set at ±5% to ±30% [50,75,76]. It is a widely recognized standard in the field of bioanalytical methods that pre-study acceptance criteria mandate that the observed mean should be within ±15% of the theoretical value, and the precision's coefficient of variation should not exceed 15% [50]. The levels of linoleic acid in buffalo milk measured using gas chromatography ranged from 51 mg/100 mL to 85.4 mg/100 mL, which is in between our quantitative ranges [8,77]. Our results show that the MIR method within the quantitative interval fully meets the above criteria.

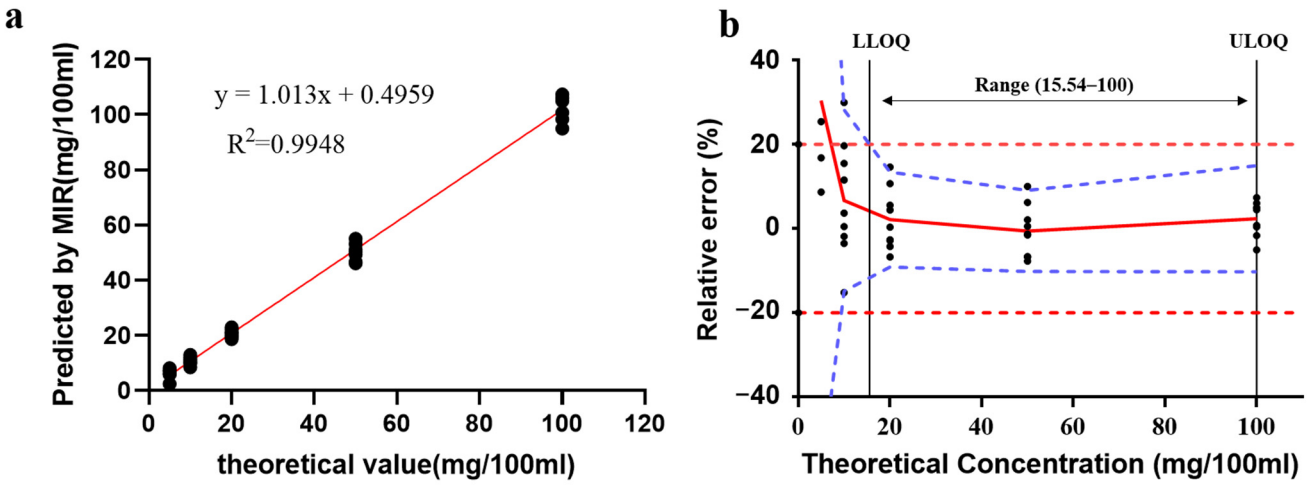

Figure 3. Linear profile and accuracy profile for MIR analysis of the linoleic acid content. (**a**) Correlation graph of MIR predictive values with reference values; (**b**) The red line represents the relative bias, the blue dashed lines are the β-expectation tolerance limits, the red dashed lines are the acceptance limits (±20%), and the 9 black points at each concentration level are relative bias for each predictive value. LLOQ represents the lower limit of quantication, and ULOQ represents the upper limit of quantication.

4. Conclusions

The objective of this study was to assess the efficacy of three machine learning models in quantifying the levels of linoleic acid (LA) in raw milk and to theoretically determine the upper and lower bounds of LA quantification. These models included partial least squares (PLSR), principal component regression (PCR), and artificial neural networks (ANNs). The study applied random forest feature selection to the models in order to improve the model performance and reduce complexity. The results of calibration and cross-validation analyses showed that the random forest partial least squares (RF-PLSR) model had the best performance among the three models, with low error values and high regression coefficients. The accuracy profile of the model was further validated using accuracy files, and it was demonstrated that Fourier transform mid-infrared (FT-MIR) could reliably quantify LA levels in the range of 15.54 mg/100 mL to 100 mg/100 mL. In conclusion, the results of this study highlight the potential of FT-MIR as a tool for rapid and reliable identification of LA content in milk. Further research efforts are recommended to develop comprehensive spectral databases for the robust assessment and reliable identification of a wider range of fatty acid concentrations. This will aid in the expansion of FT-MIR in the dairy industry and other relevant fields.

Author Contributions: Z.Y. conceived and designed the experiment data curation. X.Z. and P.N. wrote the manuscript. Z.A., C.C. and K.W. contributed animals' arrangement and sample collection. J.Z., H.L. and K.N. were responsible for software and visualization. Y.Y. and W.Z. revised the manuscript. L.Y. provided project and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 32172731) and the National Key R&D Program of China (No. 2022YFD1301001).

Data Availability Statement: The date are available from the corresponding author.

Acknowledgments: The authors thank the farmers for their cooperation in Buffalo Farm of Jinniu Animal Husbandry Co., Ltd., Hubei Province, China, Dairy Herd Improvement of Hubei detected colostrum and mature milk composition, and the authors acknowledge their contribution.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bhavadharini, B.; Dehghan, M.; Mente, A.; Rangarajan, S.; Sheridan, P.; Mohan, V.; Iqbal, R.; Gupta, R.; Lear, S.; Wentzel-Viljoen, E.; et al. Association of dairy consumption with metabolic syndrome, hypertension and diabetes in 147 812 individuals from 21 countries. *BMJ Open Diabetes Res. Care* **2020**, *8*, e000826. [CrossRef]
2. Yu, E.; Hu, F.B. Dairy Products, Dairy Fatty Acids, and the Prevention of Cardiometabolic Disease: A Review of Recent Evidence. *Curr. Atheroscler. Rep.* **2018**, *20*, 24. [CrossRef] [PubMed]
3. Chen, S.; Bobe, G.; Zimmerman, S.; Hammond, E.G.; Luhman, C.M.; Boylston, T.D.; Freeman, A.E.; Beitz, D.C. Physical and sensory properties of dairy products from cows with various milk fatty acid compositions. *J. Agric. Food Chem.* **2004**, *52*, 3422–3428. [CrossRef] [PubMed]
4. Zhiqian, L.; Rochfort, S.; Cocks, B. Milk lipidomics: What we know and what we don't. *Prog. Lipid Res.* **2018**, *71*, 70–85. [CrossRef]
5. Brown, J.M.; McIntosh, M.K. Conjugated linoleic acid in humans: Regulation of adiposity and insulin sensitivity. *J. Nutr.* **2003**, *133*, 3041–3046. [CrossRef] [PubMed]
6. Whelan, J.; Fritsche, K. Linoleic acid. *Adv. Nutr.* **2013**, *4*, 311–312. [CrossRef] [PubMed]
7. Morsy, T.A.; Kholif, A.E.; Matloup, O.H.; Elella, A.A.; Anele, U.Y.; Caton, J.S. Mustard and cumin seeds improve feed utilisation, milk production and milk fatty acids of Damascus goats. *J. Dairy Res.* **2018**, *85*, 142–151. [CrossRef] [PubMed]
8. Ferrand-Calmels, M.; Palhiere, I.; Brochard, M.; Leray, O.; Astruc, J.M.; Aurel, M.R.; Barbey, S.; Bouvier, F.; Brunschwig, P.; Caillatt, H.; et al. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. *J. Dairy Sci.* **2014**, *97*, 17–35. [CrossRef]
9. Perez-Palacios, T.; Solomando, J.C.; Ruiz-Carrascal, J.; Antequera, T. Improvements in the methodology for fatty acids analysis in meat products: One-stage transmethylation and fast-GC method. *Food Chem.* **2022**, *371*, 130995. [CrossRef]
10. Wang, F.; Chen, M.; Luo, R.; Huang, G.; Wu, X.; Zheng, N.; Zhang, Y.; Wang, J. Fatty acid profiles of milk from Holstein cows, Jersey cows, buffalos, yaks, humans, goats, camels, and donkeys based on gas chromatography–mass spectrometry. *J. Dairy Sci.* **2022**, *105*, 1687–1700. [CrossRef]
11. Sivakesava, S.; Irudayaraj, J. Rapid determination of tetracycline in milk by FT-MIR and FT-NIR Spectroscopy. *J. Dairy Sci.* **2002**, *85*, 487–493. [CrossRef]
12. Drackova, M.; Navratilova, P.; Hadra, L.; Vorlova, L.; Hudcova, L. Determination Residues of Penicillin G and Cloxacillin in Raw Cow Milk Using Fourier Transform Near Infrared Spectroscopy. *Acta Vet Brno* **2009**, *78*, 685–690. [CrossRef]
13. Grelet, C.; Bastin, C.; Gelé, M.; Davière, J.-B.; Johan, M.; Werner, A.; Reding, R.; Pierna, J.F.; Colinet, F.; Dardenne, P. Development of Fourier transform mid-infrared calibrations to predict acetone, β-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. *J. Dairy Sci.* **2016**, *99*, 4816–4825. [CrossRef]
14. Vanlierde, A.; Dehareng, F.; Gengler, N.; Froidmont, E.; McParland, S.; Kreuzer, M.; Bell, M.; Lund, P.; Martin, C.; Kuhla, B. Improving robustness and accuracy of predicted daily methane emissions of dairy cows using milk mid-infrared spectra. *J. Sci. Food Agric.* **2021**, *101*, 3394–3403. [CrossRef] [PubMed]
15. Atashi, H.; Salavati, M.; De Koster, J.; Crowe, M.; Opsomer, G.; Hostens, M. Genome-wide association for metabolic clusters in early-lactation Holstein dairy cows. *J. Dairy Sci.* **2020**, *103*, 6392–6406. [CrossRef]
16. Foldager, L.; Gaillard, C.; Sorensen, M.T.; Larsen, T.; Matthews, E.; O'flaherty, R.; Carter, F.; Crowe, M.A.; Grelet, C.; Salavati, M. Predicting physiological imbalance in Holstein dairy cows by three different sets of milk biomarkers. *Prev. Vet. Med.* **2020**, *179*, 105006. [CrossRef]
17. Tiplady, K.M.; Lopdell, T.J.; Sherlock, R.G.; Johnson, T.J.; Spelman, R.J.; Harris, B.L.; Davis, S.R.; Littlejohn, M.D.; Garrick, D.J. Comparison of the genetic characteristics of directly measured and Fourier-transform mid-infrared-predicted bovine milk fatty acids and proteins. *J. Dairy Sci.* **2022**, *105*, 9763–9791. [CrossRef]
18. Bassbasi, M.; Platikanov, S.; Tauler, R.; Oussama, A. FTIR-ATR determination of solid non fat (SNF) in raw milk using PLS and SVM chemometric methods. *Food Chem.* **2014**, *146*, 250–254. [CrossRef] [PubMed]
19. Laporte, M.F.; Paquin, P. Near-infrared analysis of fat, protein, and casein in cow's milk. *J. Agric. Food Chem.* **1999**, *47*, 2600–2605. [CrossRef]
20. Pereira, E.V.D.; Fernandes, D.D.D.; de Araujo, M.C.U.; Diniz, P.H.G.D.; Maciel, M.I.S. Simultaneous determination of goat milk adulteration with cow milk and their fat and protein contents using NIR spectroscopy and PLS algorithms. *Lwt-Food Sci. Technol.* **2020**, *127*, 109427. [CrossRef]
21. Coppa, M.; Revello-Chion, A.; Giaccone, D.; Ferlay, A.; Tabacco, E.; Borreani, G. Comparison of near and medium infrared spectroscopy to predict fatty acid composition on fresh and thawed milk. *Food Chem.* **2014**, *150*, 49–57. [CrossRef]
22. Flores-Valdez, M.; Meza-Márquez, O.G.; Osorio-Revilla, G.; Gallardo-Velázquez, T. Identification and Quantification of Adulterants in Coffee (*Coffea arabica* L.) Using FT-MIR Spectrosc. *Coupled Chemom. Foods* **2020**, *9*, 851.
23. El Orche, A.; Mamad, A.; Elhamdaoui, O.; Cheikh, A.; El Karbane, M.; Bouatia, M. Comparison of Machine Learning Classification Methods for Determining the Geographical Origin of Raw Milk Using Vibrational Spectroscopy. *J. Spectrosc.* **2021**, *2021*, 5845422. [CrossRef]
24. Amsaraj, R.; Ambade, N.D.; Mutturi, S. Variable selection coupled to PLS2, ANN and SVM for simultaneous detection of multiple adulterants in milk using spectral data. *Int. Dairy J.* **2021**, *123*, 105172. [CrossRef]
25. Lovatti, B.P.; Nascimento, M.H.; Neto, Á.C.; Castro, E.V.; Filgueiras, P.R. Use of Random forest in the identification of important variables. *Microchem. J.* **2019**, *145*, 1129–1134. [CrossRef]

26. Wu, X.-M.; Zhang, Q.-Z.; Wang, Y.-Z. Traceability of wild Paris polyphylla Smith var. yunnanensis based on data fusion strategy of FT-MIR and UV–Vis combined with SVM and random forest. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2018**, *205*, 479–488. [CrossRef]
27. Beriain, M.J.; Ibanez, F.C.; Beruete, E.; Gomez, I.; Beruete, M. Estimation of Fatty Acids in Intramuscular Fat of Beef by FT-MIR Spectroscopy. *Foods* **2021**, *10*, 155. [CrossRef]
28. Bonfatti, V.; Degano, L.; Menegoz, A.; Carnier, P. Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. *J. Dairy Sci.* **2016**, *99*, 8216–8221. [CrossRef] [PubMed]
29. Mossoba, M.; Kramer, J.; Fritsche, J.; Yurawecz, M.; Eulitz, K.; Ku, Y.; Rader, J. Application of standard addition to eliminate conjugated linoleic acid and other interferences in the determination of total trans fatty acids in selected food products by infrared spectroscopy. *J. Am. Oil Chem. Soc.* **2001**, *78*, 631–634. [CrossRef]
30. Caredda, M.; Addis, M.; Ibba, I.; Leardi, R.; Scintu, M.F.; Piredda, G.; Sanna, G. Prediction of fatty acid content in sheep milk by Mid-Infrared spectrometry with a selection of wavelengths by Genetic Algorithms. *Lwt-Food Sci. Technol.* **2016**, *65*, 503–510. [CrossRef]
31. Hemmateenejad, B.; Yousefinejad, S. Multivariate standard addition method solved by net analyte signal calculation and rank annihilation factor analysis. *Anal. Bioanal. Chem.* **2009**, *394*, 1965–1975. [CrossRef]
32. Zhang, T.-M.; Yuan, B.; Liang, Y.-Z.; Cao, J.; Pan, C.-X.; Ying, B.; Lu, D.-Y. Elimination of matrix effect and simultaneous determination of multi-components in complex systems by matrix coefficient non-linearity multivariate calibration based on single point response signals. *Anal. Sci.* **2007**, *23*, 581–587. [CrossRef] [PubMed]
33. Barbas, K.H.; O'Brien, K.; Forbes, P.W.; Belfort, M.B.; Connor, J.A.; Thiagarajan, R.R.; Huh, S.Y. Macronutrient Analysis of Modified-Fat Breast Milk Produced by 3 Methods of Fat Removal. *J. Parenter. Enter. Nutr.* **2020**, *44*, 895–902. [CrossRef] [PubMed]
34. Yurchenko, S.; Sats, A.; Poikalainen, V.; Karus, A. Method for determination of fatty acids in bovine colostrum using GC-FID. *Food Chem.* **2016**, *212*, 117–122. [CrossRef] [PubMed]
35. Soyeurt, H.; Grelet, C.; McParland, S.; Calmels, M.; Coffey, M.; Tedde, A.; Delhez, P.; Dehareng, F.; Gengler, N. A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra. *J. Dairy Sci.* **2020**, *103*, 11585–11596. [CrossRef] [PubMed]
36. Belay, T.K.; Dagnachew, B.S.; Kowalski, Z.M.; Ådnøy, T. An attempt at predicting blood β-hydroxybutyrate from Fourier-transform mid-infrared spectra of milk using multivariate mixed models in Polish dairy cattle. *J. Dairy Sci.* **2017**, *100*, 6312–6326. [CrossRef] [PubMed]
37. Portnoy, M.; Coon, C.; Barbano, D. Infrared milk analyzers: Milk urea nitrogen calibration. *J. Dairy Sci.* **2021**, *104*, 7426–7437. [CrossRef]
38. Tibble, H.; Chan, A.; Mitchell, E.A.; Horne, E.; Doudesis, D.; Horne, R.; Mizani, M.A.; Sheikh, A.; Tsanas, A. A data-driven typology of asthma medication adherence using cluster analysis. *Sci. Rep.* **2020**, *10*, 14999. [CrossRef]
39. Smith, A.E.; Mason, A.K. Cost estimation predictive modeling: Regression versus neural network. *Eng. Econ.* **1997**, *42*, 137–161. [CrossRef]
40. Wang, N.; Er, M.J.; Han, M. Generalized single-hidden layer feedforward networks for regression problems. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *26*, 1161–1176. [CrossRef]
41. Dreiseitl, S.; Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. *J. Biomed. Inform.* **2002**, *35*, 352–359. [CrossRef]
42. Rogers, J.; Gunn, S. Identifying feature relevance using a random forest. In *Subspace, Latent Structure and Feature Selection*; Saunders, G., Grobelnik, M., Gunn, S., ShaweTaylor, J., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3940, pp. 173–184.
43. Archer, E.; Archer, M.E. *Package 'rfPermute'*; R Project: Indianapolis, IN, USA, 2016.
44. Miranda, E.N.; Barbosa, B.H.G.; Silva, S.H.G.; Monti, C.A.U.; Tng, D.Y.P.; Gomide, L.R. Variable selection for estimating individual tree height using genetic algorithm and random forest. *For. Ecol. Manag.* **2022**, *504*, 119828. [CrossRef]
45. Liu, J.; Wen, Y.; Dong, N.; Lai, C.L.; Zhao, G.H. Authentication of lotus root powder adulterated with potato starch and/or sweet potato starch using Fourier transform mid-infrared spectroscopy. *Food Chem.* **2013**, *141*, 3103–3109. [CrossRef]
46. Xue, Z.; Xu, B.; Yang, C.; Cui, X.L.; Li, J.Y.; Shi, X.Y.; Qiao, Y.J. Method validation for the analysis of licorice acid in the blending process by near infrared diffuse reflectance spectroscopy. *Anal. Methods* **2015**, *7*, 5830–5837. [CrossRef]
47. Rozet, E.; Ceccato, A.; Hubert, C.; Ziemons, E.; Oprean, R.; Rudaz, S.; Boulanger, B.; Hubert, P. Analysis of recent pharmaceutical regulatory documents on analytical method validation. *J Chromatogr A* **2007**, *1158*, 111–125. [CrossRef] [PubMed]
48. Food and Drug Administration. Bioanalytical Method Validation Guidance for Industry. 2018. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/bioanalytical-method-validation-guidance-industry (accessed on 1 February 2023)
49. Hubert, P.; Chiap, P.; Crommen, J.; Boulanger, B.; Chapuzet, E.; Mercier, N.; Bervoas-Martin, S.; Chevalier, P.; Grandjean, D.; Lagorce, P. The SFSTP guide on the validation of chromatographic methods for drug bioanalysis: From the Washington Conference to the laboratory. *Anal. Chim. Acta* **1999**, *391*, 135–148. [CrossRef]
50. Hoffman, D.; Kringle, R. A total error approach for the validation of quantitative analytical methods. *Pharm Res* **2007**, *24*, 1157–1164. [CrossRef] [PubMed]

51. Kulkarni, P.; Kushary, D. β Expectation and β-content tolerance intervals for dependent observations. *Commun. Stat. Theory Methods* **1991**, *20*, 1043–1054. [CrossRef]
52. Saffaj, T.; Ihssane, B.; Jhilal, F.; Bouchafra, H.; Laslami, S.; Sosse, S.A. An overall uncertainty approach for the validation of analytical separation methods. *Analyst* **2013**, *138*, 4677–4691. [CrossRef] [PubMed]
53. Saffaj, T.; Ihssane, B. Uncertainty profiles for the validation of analytical methods. *Talanta* **2011**, *85*, 1535–1542. [CrossRef]
54. Wang, Q.Y.; Bovenhuis, H. Validation strategy can result in an overoptimistic view of the ability of milk infrared spectra to predict methane emission of dairy cattle. *J. Dairy Sci.* **2019**, *102*, 6288–6295. [CrossRef]
55. Tao, F.F.; Ngadi, M. Applications of spectroscopic techniques for fat and fatty acids analysis of dairy foods. *Curr Opin Food Sci* **2017**, *17*, 100–112. [CrossRef]
56. Du, C.; Wei, J.; Wang, S.; Jia, Z. Genomic selection using principal component regression. *Heredity* **2018**, *121*, 12–23. [CrossRef] [PubMed]
57. Desta, F.; Buxton, M.; Jansen, J. Fusion of mid-wave infrared and long-wave infrared reflectance spectra for quantitative analysis of minerals. *Sensors* **2020**, *20*, 1472. [CrossRef] [PubMed]
58. Wang, C.; Qu, L.; Yang, L.; Liu, D.; Morrissey, E.; Miao, R.; Liu, Z.; Wang, Q.; Fang, Y.; Bai, E. Large-scale importance of microbial carbon use efficiency and necromass to soil organic carbon. *Glob. Chang. Biol.* **2021**, *27*, 2039–2048. [CrossRef]
59. Chen, W.; Xu, R.; Chen, J.; Yuan, X.; Zhou, L.; Tan, T.; Fan, J.; Zhang, Y.; Hu, T. Consistent responses of surface-and subsurface soil fungal diversity to N enrichment are mediated differently by acidification and plant community in a semi-arid grassland. *Soil Biol. Biochem.* **2018**, *127*, 110–119. [CrossRef]
60. Oehm, A.W.; Springer, A.; Jordan, D.; Strube, C.; Knubben-Schweizer, G.; Jensen, K.C.; Zablotski, Y. A machine learning approach using partitioning around medoids clustering and random forest classification to model groups of farms in regard to production parameters and bulk tank milk antibody status of two major internal parasites in dairy cows. *PLoS ONE* **2022**, *17*, e0271413. [CrossRef] [PubMed]
61. Szymczak, S.; Holzinger, E.; Dasgupta, A.; Malley, J.D.; Molloy, A.M.; Mills, J.L.; Brody, L.C.; Stambolian, D.; Bailey-Wilson, J.E. r2VIM: A new variable selection method for random forests in genome-wide association studies. *Biodata Min.* **2016**, *9*, 7. [CrossRef]
62. Cammarota, C.; Pinto, A. Variable selection and importance in presence of high collinearity: An application to the prediction of lean body mass from multi-frequency bioelectrical impedance. *J. Appl. Stat.* **2021**, *48*, 1644–1658. [CrossRef]
63. Verikas, A.; Gelzinis, A.; Bacauskiene, M. Mining data with random forests: A survey and results of new tests. *Pattern Recogn.* **2011**, *44*, 330–349. [CrossRef]
64. Christy, A.A.; Egeberg, P.K. Quantitative determination of saturated and unsaturated fatty acids in edible oils by infrared spectroscopy and chemometrics. *Chemom. Intell Lab.* **2006**, *82*, 130–136. [CrossRef]
65. El Jabri, M.; Trossat, P.; Wolf, V.; Beuvier, E.; Rolet-Répécaud, O.; Gavoye, S.; Gaüzère, Y.; Belysheva, O.; Gaudillière, N.; Notz, E. Mid-infrared spectrometry prediction of the cheese-making properties of raw Montbéliarde milks from herds and cheese dairy vats used for the production of Protected Designation of Origin and Protected Geographical Indication cheeses in Franche-Comté. *J. Dairy Sci.* **2020**, *103*, 5992–6002. [CrossRef]
66. Sanchez, M.P.; El Jabri, M.; Minéry, S.; Wolf, V.; Beuvier, E.; Laithier, C.; Delacroix-Buchet, A.; Brochard, M.; Boichard, D. Genetic parameters for cheese-making properties and milk composition predicted from mid-infrared spectra in a large data set of Montbéliarde cows. *J. Dairy Sci.* **2018**, *101*, 10048–10061. [CrossRef]
67. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin, Germany, 1999.
68. Mota, L.F.M.; Giannuzzi, D.; Bisutti, V.; Pegolo, S.; Trevisi, E.; Schiavon, S.; Gallo, L.; Fineboym, D.; Katz, G.; Cecchinato, A. Real-time milk analysis integrated with stacking ensemble learning as a tool for the daily prediction of cheese-making traits in Holstein cattle. *J. Dairy Sci.* **2022**, *105*, 4237–4255. [CrossRef] [PubMed]
69. Martin, M.J.; Dórea, J.; Borchers, M.; Wallace, R.; Bertics, S.; DeNise, S.; Weigel, K.; White, H. Comparison of methods to predict feed intake and residual feed intake using behavioral and metabolite data in addition to classical performance variables. *J. Dairy Sci.* **2021**, *104*, 8765–8782. [CrossRef] [PubMed]
70. Rutten, M.; Bovenhuis, H.; Hettinga, K.A.; Valenberg, H.; Arendonk, J. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. *J. Dairy Sci.* **2009**, *92*, 6202–6209. [CrossRef]
71. Xue, Z.; Xu, B.; Liu, Q.; Shi, X.-Y.; Li, J.-Y.; Wu, Z.-S.; Qiao, Y.-J. Application of uncertainty assessment in NIR quantitative analysis of Traditional Chinese Medicine. *Guang Pu Xue Yu Guang Pu Fen Xi* **2014**, *34*, 2657–2661.
72. Borman, P.; Chatfield, M.; Asahara, H.; Tamura, F.; Watkins, A. Risk-Based Intermediate Precision Studies for Analytical Procedure Validation. *Pharm. Technol. Regul. Sourceb* **2019**, *2019*, 12–22.
73. Shewiyo, D.H.; Kaale, E.; Risha, P.; Dejaegher, B.; De Beer, J.; Smeyers-Verbeke, J.; Vander Heyden, Y. Accuracy profiles assessing the validity for routine use of high-performance thin-layer chromatographic assays for drug formulations. *J. Chromatogr. A* **2013**, *1293*, 159–169. [CrossRef]
74. Schaefer, C.; Clicq, D.; Lecomte, C.; Merschaert, A.; Norrant, E.; Fotiadu, F. A Process Analytical Technology (PAT) approach to control a new API manufacturing process: Development, validation and implementation. *Talanta* **2014**, *120*, 114–125. [CrossRef] [PubMed]
75. Frampas, C.; Ney, J.; Coburn, M.; Augsburger, M.; Varlet, V. Xenon detection in human blood: Analytical validation by accuracy profile and identification of critical storage parameters. *J. Forensic Leg. Med.* **2018**, *58*, 14–19. [CrossRef] [PubMed]

76. Almeida, J.; Bezerra, M.; Markl, D.; Berghaus, A.; Borman, P.; Schlindwein, W. Development and Validation of an in-line API Quantification Method Using AQbD Principles Based on UV-Vis Spectroscopy to Monitor and Optimise Continuous Hot Melt Extrusion Process. *Pharmaceutics* **2020**, *12*, 150. [CrossRef] [PubMed]
77. Teng, F.; Wang, P.; Yang, L.; Ma, Y.; Day, L. Quantification of Fatty Acids in Human, Cow, Buffalo, Goat, Yak, and Camel Milk Using an Improved One-Step GC-FID Method. *Food Anal. Method* **2017**, *10*, 2881–2891. [CrossRef]