foods

Article

Several Feature Extraction Methods Combined with Near-Infrared Spectroscopy for Identifying the Geographical Origins of Milk

Xiaohong Wu [1,2,*], Yixuan Wang [1], Chengyu He [1], Bin Wu [3,*], Tingfei Zhang [1] and Jun Sun [1]

[1] School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China; 2222207041@stmail.ujs.edu.cn (Y.W.); 2222107003@stmail.ujs.edu.cn (C.H.); 2221907104@stmail.ujs.edu.cn (T.Z.); sun2000jun@ujs.edu.cn (J.S.)
[2] High-Tech Key Laboratory of Agricultural Equipment and Intelligence of Jiangsu Province, Jiangsu University, Zhenjiang 212013, China
[3] Department of Information Engineering, Chuzhou Polytechnic, Chuzhou 239000, China
* Correspondence: wxh419@ujs.edu.cn (X.W.); wubin@chzc.edu.cn (B.W.)

Abstract: Milk is a kind of dairy product with high nutritive value. Tracing the origin of milk can uphold the interests of consumers as well as the stability of the dairy market. In this study, a fuzzy direct linear discriminant analysis (FDLDA) is proposed to extract the near-infrared spectral information of milk by combining fuzzy set theory with direct linear discriminant analysis (DLDA). First, spectral data of the milk samples were collected by a portable NIR spectrometer. Then, the data were preprocessed by Savitzky–Golay (SG) and standard normal variables (SNV) to reduce noise, and the dimensionality of the spectral data was decreased by principal component analysis (PCA). Furthermore, linear discriminant analysis (LDA), DLDA, and FDLDA were employed to transform the spectral data into feature space. Finally, the k-nearest neighbor (KNN) classifier, extreme learning machine (ELM) and naïve Bayes classifier were used for classification. The results of the study showed that the classification accuracy of FDLDA was higher than DLDA when the KNN classifier was used. The highest recognition accuracy of FDLDA, DLDA, and LDA could reach 97.33%, 94.67%, and 94.67%. The classification accuracy of FDLDA was also higher than DLDA when using ELM and naïve Bayes classifiers, but the highest recognition accuracy was 88.24% and 92.00%, respectively. Therefore, the KNN classifier outperformed the ELM and naïve Bayes classifiers. This study demonstrated that combining FDLDA, DLDA, and LDA with NIR spectroscopy as an effective method for determining the origin of milk.

Keywords: milk; near-infrared spectroscopy; feature extraction; geographical origins; classification

check for updates

Citation: Wu, X.; Wang, Y.; He, C.; Wu, B.; Zhang, T.; Sun, J. Several Feature Extraction Methods Combined with Near-Infrared Spectroscopy for Identifying the Geographical Origins of Milk. *Foods* **2024**, *13*, 1783. https://doi.org/10.3390/foods13111783

Academic Editors: Manuel Castillo Zambudio and Thierry Aussenac

Received: 2 April 2024
Revised: 17 May 2024
Accepted: 4 June 2024
Published: 6 June 2024

1. Introduction

Milk is a natural dairy product with excellent nutritional value and is popular with people around the world [1]. The content of nutrients in milk is related to climate, original pasture environment, animal feed, and other factors [2]. The origin of milk is an important element in the quality of milk and influences the price of the product. In Europe, there are policies such as the protected designation of origin (PDO) or protected geographical indication (PGI) to protect local production [3]. Due to the protected designation of origin (PDO) policies, products with PDO status command higher market values, which unfortunately makes them a prime target for counterfeiting, particularly through the forgery of product labels. Incorrect brand labels not only do harm to the reputation of the origin but also infringe on the legitimate rights and interests of consumers [4]. The traceability of milk origin is not only crucial for ensuring product safety and quality, but it also plays a significant role in safeguarding brands and promoting specialty products in their respective source regions. Therefore, finding a quick way to determine the geographical origin of milk is essential for milk food safety concerns and brand protection.

There are some methods for identifying the geographical origin of food products [5–8]. Among them, the use of stable isotope and elemental analysis is a common approach [9,10]. The distribution of isotopes and element contents in food items is affected by many factors, especially the location of origin [11]. Ng et al. [12] accurately differentiated the geographical origin of 307 milk samples by analyzing their isotopic fingerprints and elemental profiles. In addition, multiple chromatographic approaches have been developed to successfully detect the geographical origin of food products [13]. In conjunction with the nutritional composition and geographical features of milk, Xie et al. [14] employed high-performance liquid chromatography (HPLC), elemental analyzer isotope ratio mass spectrometry (EA-IRMS), and inductively coupled plasma mass spectrometry (ICP-MS) to establish the traceability of milk samples from Neimenggu. Moreover, the proton transfer reaction mass spectrometry (PTR-MS) technology is another effective method [15]. Although the above several methods can be applied to obtain the geographical information of food, they are often complex, expensive, destructive, and unsuitable for widespread use.

NIR spectroscopy is an emerging non-destructive analysis technology with the advantages of being fast, simple, and low-cost, and it has been widely applied for the analysis and classification of food [16–18]. Previous studies provide numerous examples where NIR spectroscopy has been utilized to facilitate both the classification of milk and the traceability of its origin. Infrared light measures the vibrations of molecules. Each functional group or structural feature of a molecule has a unique vibrational frequency [19]. When the effects of all the different functional groups are put together, a unique molecular "fingerprint" is formed that can be used to identify the sample [19]. Therefore, infrared spectroscopy can determine the composition of milk very quickly for qualitative analysis of milk.

Employing a synergistic approach, Zhang et al. [20] swiftly distinguished adulterated milk samples, which contained various pseudo-proteins and thickeners, from authentic raw milk samples. This achievement was made possible through the integration of near-infrared (NIR) spectroscopy with an advanced support vector machine (SVM) and an enhanced, non-linear simplified k-nearest neighbors (KNN) identification model. Diaz-Olivares et al. developed an on-farm online milk composition analysis system that employed NIR spectroscopy to analyze and model the composition of milk [21]. Pereira et al. successfully used NIR spectroscopy and the PLS algorithm to classify cow milk, goat milk, and mixed milk, demonstrating that NIR spectroscopy was an excellent method for the analysis of adulteration in milk [22]. Coppa et al. utilized NIR spectroscopy to acquire the feature information of milk samples to accurately distinguish pastured milk from non-pastured milk and trace their geographical origin [23]. Yin used infrared spectroscopy to test milk from four origins and developed a regression model that not only allowed for a good determination of the origin of the milk but also allowed for the detection of adulterated milk with different concentrations of urea and glucose [24].

Commercial spectroscopic equipment is often bulky and expensive, and can usually only be used in a laboratory environment, making it difficult to meet the requirement for fast, accurate, and real-time analysis in some conditions [25]. In recent years, with the advancement of science and technology, as well as the growing demand for practical applications, portable NIR spectrometers have been widely used for the analysis of samples [26]. The portable NIR spectrometers have been made more convenient for their small size, but their spectral resolution is lower than that of bulky commercial NIR spectrometers, and mathematical modeling is necessary to alleviate this weakness. The NIR spectral data of milk recorded by the portable NIR spectrometer are biased, and there are some overlapped data points between different categories of spectral data, which will have a significant impact on the model's classification accuracy [27]. The feature extraction approach can extract meaningful information from the obtained spectral data to increase the model's classification accuracy.

When dealing with high-dimensional data, the classical linear discriminant analysis (LDA) algorithm has several drawbacks [28]. For example, if the dimensionality of data is much greater than the number of samples, this can lead to the singularity of scattering matrices, making it difficult to find the optimal projection direction. This does not satisfy the requirement of the LDA algorithm for the within-class scattering matrix. Direct linear discriminant analysis (DLDA) is an improved dimensionality reduction technique, which extends the application of the classical LDA algorithm in the case of small sample data [29,30]. The DLDA algorithm is implemented by diagonalizing the between-class scattering matrix S_b and the within-class scattering matrix S_w at the same time. It discards the null space of the between-class scattering matrix S_b, but preserves the null space of the within-class scattering matrix S_w that contains the most discriminant information. Generally, compared with LDA, the discriminant vectors of DLDA often have less redundant information and can achieve better classification results. However, DLDA is still unable to extract useful information from the overlapped NIR spectra. The theory of fuzzy set is an effective method to cope with unclear data in the field of pattern recognition. Patel et al. used the fuzzy k-nearest neighbor method to classify data instances into distinct groups. By considering both the nearest neighbors and their distances from the instances, this method effectively minimized the risk of classification errors [31]. Wang brought two fuzzy notions into LDA and successfully built a fuzzy face recognition model by investigating the structure of each layer of the two types of fuzzy systems [32]. The core thought of fuzzy set theory is to search for a parameter that measures the extent to which a sample belongs to a particular category, and this parameter is known as fuzzy membership. By recognizing the objective existence of fuzzy items, fuzzy set theory adeptly addresses the challenges posed by uncertain concepts This makes it particularly useful for extracting features from overlapped spectral data [27].

This study prefers qualitative analyses of milk from different origins. The aim is to differentiate between milk from different origins using a method that is as quick, simple, low-cost, and non-destructive as possible, whereas conducting laboratory experiments would not only take a long time and cost but also cause damage to the milk samples, which is not in line with the fundamental aim of this study. Therefore, no laboratory measurements were required in this study.

The purpose of this study was to create a model that could rapidly and reliably identify the geographical origin of milk by combining fuzzy set theory with feature extraction methods. The research steps were as follows: (1) A portable NIR spectrometer was employed to collect spectral data from milk samples. (2) Several data processing algorithms were integrated to deal with the spectra. Savitzky–Golay (SG) and standard normal variables (SNV) were conducted to remove noise, and principal component analysis (PCA) was applied for dimension reduction. Three feature extraction methods were used to separate the spectral data. (3) The k-nearest neighbor (KNN) classifier, extreme learning machine (ELM), and naïve Bayes classifiers were used to classify the data after feature extraction.

2. Materials and Experiments

2.1. Internal Content Difference between Milks

The nutrient content in milk is related to the place of origin [33]. Ningxia had favorable temperatures all year round, which provided sufficient feed resources for dairy farming [34]. While Henan dairy cattle feed mainly relied on straw and natural pasture, there was a serious shortage of high-quality feed. Table 1 summarizes the nutrient content of milk from different regions.

Table 1. Comparison of the nutrient content of milk in different regions.

Origin	Fat (%)	Protein (%)	Lactose (%)	Total Solids (%)
Henan	3.18 ± 1.16^{c}	3.51 ± 0.47^{b}	4.95 ± 0.23^{b}	12.14 ± 1.38^{c}
Neimenggu	3.44 ± 0.92^{b}	3.43 ± 0.43^{b}	5.09 ± 0.23^{b}	12.45 ± 1.21^{b}
Ningxia	3.52 ± 0.85^{b}	3.67 ± 0.41^{a}	5.06 ± 0.24^{a}	12.78 ± 0.18^{ab}

Values in the table are mean ± standard deviation, different letters labelled on the shoulder of peer data indicate significant differences ($p < 0.05$), and having the same letter indicates non-significant differences ($p > 0.05$).

The origin of milk could be distinguished by a stable isotope [35]. The $\delta^{13}C$ of milk was mainly related to the composition of the cow's food [36]. $\delta^{15}N$ mainly reflected the soil conditions of the environment in which the cow lived and the number of legumes in the ration. $\delta^{18}O$ and $\delta^{2}H$ were closely related to the environmental factors of the local water source, air, and climatic conditions. Table 2 shows the results of stable isotope ratio measurements of milk samples from different regions.

Table 2. Results of stable isotope ratio measurements of milk samples from different regions.

Origin	$\delta^{13}C$	$\delta^{15}N$	$\delta^{18}O$	$\delta^{2}H$
Heilongjiang	−18.389 ± 0.252 a	2.976 ± 0.193 b	−158.990 ± 15.181 c	8.858 ± 1.807 bc
Hebei	−18.456 ± 0.139 ab	3.309 ± 0.075 a	−144.723 ± 2.129 b	9.880 ± 0.297 b
Neimenggu	−19.138 ± 0.429 c	3.501 ± 0.139 a	−147.340 ± 11.573 bc	10.036 ± 0.350 b

Values in the table are mean ± standard deviation, different letters labelled on the shoulder of peer data indicate significant differences (p < 0.05), and having the same letter indicates non-significant differences (p > 0.05).

Differences in mineral elements also provided important support for identifying the geographical origin of milk [37]. Table 3 shows the contents and significant differences of different mineral elements in milk samples. These differences were caused by the different feeding methods of the cows. The trace elements were present at low levels in living organisms and depended mainly on soil and geological features.

Table 3. The mineral content of milk from different origins.

Element	Hebei	Neimenggu	Ningxia
Na (mg/kg)	2046 ± 429 ^{a}	2695 ± 422 ^{b}	3358 ± 511 ^{c}
Mg (mg/kg)	879 ± 129 ^{a}	1232 ± 134 ^{b}	1161 ± 363 ^{c}
K (mg/kg)	12,899 ± 2773 ^{a}	156,616 ± 2179 ^{b}	22,872 ± 5023 ^{c}
Ca (mg/kg)	8113 ± 1073 ^{b}	3843 ± 487 ^{a}	12,042 ± 3399 ^{c}
Ti (mg/kg)	40.5 ± 7.80 ^{a}	64.8 ± 7.20 ^{d}	56.5 ± 9.39 ^{b}
Cr (mg/kg)	44.0 ± 7.71 ^{a}	176 ± 29.5 ^{d}	88.3 ± 18.3 ^{b}
Mn (mg/kg)	460 ± 130 ^{a}	1055 ± 216 ^{c}	751 ± 256 ^{b}
Fe (mg/kg)	62.9 ± 9.63 ^{a}	70.2 ± 7.66 ^{b}	84.6 ± 13.3 ^{c}
Ni (mg/kg)	348 ± 53.3 ^{a}	549 ± 82.0 ^{bc}	585 ± 130 ^{c}
Zn (mg/kg)	34.9 ± 7.06 ^{a}	75.3 ± 12.0 ^{c}	34.3 ± 7.84 ^{a}
Sr (mg/kg)	4.10 ± 0.935 ^{a}	11.5 ± 2.16 ^{c}	7.96 ± 1.49 ^{b}

Values in the table are mean ± standard deviation, different letters labelled on the shoulder of peer data indicate significant differences (p < 0.05), and having the same letter indicates non-significant differences (p > 0.05).

Table 3 shows that there are differences in the type and content of components in milk from different origins. Therefore, the use of NIR spectroscopy and feature extraction techniques to classify the origin of milk is scientifically sound and also highly feasible.

2.2. Sample Preparation

The five different commercial milks used in the experiment were purchased from the local supermarket in Zhenjiang, China. They came from five different provinces: Heilongjiang, Hebei, Henan, Ningxia, and Neimenggu, which are the main dairy production bases in China. Samples needed to be selected with attention to the exact origin information, date of manufacture, and nutritional content labeled on the box. To ensure the objectivity of the experiment, all five different commercial milks were purchased on the same day and the production date of them needed to be as close as possible. To test the classification performance of the model, semi-skimmed milk was selected for milk from Hebei, whole milk was selected for milk from the other four origins. Ultra-high temperature sterilized milk was selected for all milk samples to facilitate subsequent experiments and storage. For a total of 300 samples, 60 of each type of milk were obtained. All milk samples were kept in the laboratory at a temperature of 25 ± 1 °C to preserve the milk's quality.

2.3. Instruments and Software

A reflectance handheld NIR spectrometer NIR-M-R2 (Shenzhen Pynect Science and Technology Co., Ltd., Shenzhen, China) was used in this study. The spectrometer captured spectra in the wavelength range of 900–1700 nm with an optical resolution of 10 nm, a signal-to-noise ratio of 6000:1, and a slit size of 1.8 mm × 0.025 mm. It also had a temperature and humidity sensor for displaying temperature and humidity measurements in the module.

All algorithms mentioned were programmed and executed in MATLAB (The Mathworks Inc., Natick, MA, USA) 2019b on a personal computer with the operating system Microsoft Windows 10.

2.4. Spectra Collection

Firstly, the spectral dimension in this experiment was set to 228 and the spectrometer was turned on and preheated for half an hour. After that, 5 mL of milk was taken from each milk sample by using a pipette and placed in a 10 mL beaker. During the scanning process, the beaker needed to be kept at a constant distance from the portable NIR spectrometer to ensure the stability of the experiment. The milk NIR spectra recorded in this experiment had a wavelength range of 5882 cm^{-1}–11,111 cm^{-1} and a data dimension of 228. Each milk sample was used three times, and the average value of the three spectral data served as the final NIR spectral data.

2.5. NIR Spectra Preprocessing

When collecting spectral data, due to the impacts of measurement errors and optical scattering, the raw spectral data were mixed with a large amount of noise. To improve the accuracy of the classification, the raw spectra must be preprocessed to remove the influence of noise and scattering on the spectral data [38].

Multiplying scattering correction (MSC), standard normal variables (SNV), Savitzky–Golay (SG) and mean centering (MC) were used in this study. For these four data preprocessing methods, each serves a distinct purpose [39]. MSC is primarily employed to minimize the scattering effect induced by irregular particle distribution and particle size. SNV is applied to remove the impact of variations in solid particle size, apparent scattering, and optical paths. SG can remove scattering phenomena, increase spectral data smoothness, and lower the impact of diffuse reflections and noise interference. The MC could magnify weak signals.

2.6. Division of Training Set and Test Set

A total of 300 spectral data were obtained. The data were divided into a training set and a test set in a ratio of 3:1, with 45 training samples and 15 test samples for each milk sample. This means that there were 225 samples in the training set and 75 samples in the test set.

2.7. Principal Component Analysis

PCA is the most commonly used dimensionality reduction algorithm in the field of machine learning. PCA can map the features of spectral data into a few important features while discarding unnecessary features. PCA is always utilized to reduce data dimension and remove the redundant information in the spectral data. It reduces the dimensionality of data and maintains the most important information of the high-dimensional data by orthogonally transforming a large number of significant variables into a few uncorrelated groups of variables. PCA can remove noise and unnecessary features, and this is advantageous to improve data processing speed and classification accuracy. In practical application, the amount of the chosen principal components has a significant impact on the classification results [40,41]. Selecting appropriate principal components and retaining the relevant important features are conducive to the improvement of classification accuracy.

3. Feature Extraction Methods

NIR spectroscopy suffers from defects such as wide spectral band, severe overlap, and complex information resolution, so it can only be used as an indirect measurement method. Therefore, appropriate metrological models need to be established to qualitatively analyze the samples. Feature extraction converts the target sample set from a high-dimensional feature space to a low-dimensional feature space, making the projected sample easy to distinguish.

3.1. Linear Discriminant Analysis

The LDA algorithm is a supervised feature extraction method. Using a given training set of samples, the sample points are projected onto a straight line. To obtain the minimum distances within a class and the maximum distances between classes, the projection points of samples belonging to the same class need to be as close together as feasible, while those belonging to different classes need to be as far apart. However, when the dimension of the sample data is larger than the number of sample data, it will result in the required scattering matrix being a singular matrix, which prevents the subsequent calculations. Therefore, an improvement of LDA is needed.

3.2. Direct Linear Discriminant Analysis

Direct linear discriminant analysis (DLDA) is an improved feature extraction method that extends the application of the classical LDA algorithm in the context of small sample setting (SSS). At the heart of DLDA is the idea of simultaneous diagonalization, which tries to find a matrix that can directly diagonalize the scattering matrix [42]. DLDA is computed in two phases: the first phase computes the conversion matrix to transform the training samples into the range space of the scattering matrix; the second phase further reduces the dimensionality of the samples through some conditioning matrices [30]. Because the NIR spectral data of milk have high dimensions, DLDA can be utilized to analyze the data and extract useful information.

The specific steps of the DLDA algorithm are as follows (input: data matrix X; output: transformation matrix W_{DLDA}) [42]:

- Step 1. Calculate matrix S_w and S_b;
- Step 2. Apply the eigendecomposition on S_b to get Y and D_b, then $Z \leftarrow YD_b^{-1/2}$;
- Step 3. Compute the eigenvalues and eigenvectors of $Z^T S_w Z$ to get D_w and U;
- Step 4. Obtain the transformation matrix: $W_{DLDA} = D_w^{-1/2} U^T Z^T$.

In step 1, the within-class scattering matrix S_w and the between-class scattering matrix S_b are calculated as follows:

$$S_w = \sum_{i=1}^{c} \sum_{k=1}^{n_i} (\overline{x}_i - x_k)(\overline{x}_i - x_k)^T \tag{1}$$

$$S_b = \sum_{i=1}^{c} n_i (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})^T \tag{2}$$

In the above equations, c represents the number of sample categories; n_i represents the number of samples in the ith category; $\overline{x}$ represents the average value of the total sample; and $\overline{x}_i$ represents the average value of the ith category.

3.3. Fuzzy Direct Linear Discriminant Analysis

When faced with complex datasets, DLDA sometimes fails to extract feature information from data, and it is especially difficult to solve the problem of overlapped data. Fuzzy set theory is an effective method for analyzing uncertain data.

Before using the fuzzy theory, the initial cluster centers and fuzzy membership values need to be calculated first. The initial cluster center of fuzzy direct linear discriminant analysis (FDLDA) was the mean value of each kind of training sample. The parameters were set as follows: the fuzzy weight coefficient $m = 3.5$, the number of sample category $c = 5$, and the five initial cluster centers constituted the matrix $V^{(0)}$ [27]:

$$V^{(0)} = \begin{bmatrix} v_1^{(0)} \\ v_2^{(0)} \\ v_3^{(0)} \\ v_4^{(0)} \\ v_5^{(0)} \end{bmatrix} = \begin{bmatrix} -0.5798 & -0.0148 & 0.0123 & 0.1106 & 0.1268 \\ -0.1045 & 0.1909 & 0.0445 & -0.0329 & -0.0006 \\ 0.1359 & 0.3192 & -0.0618 & -0.0250 & -0.0645 \\ 0.8331 & -0.1365 & -0.0003 & 0.0707 & 0.0170 \\ -0.2921 & -0.3019 & -0.0264 & -0.0652 & -0.1108 \end{bmatrix} \tag{3}$$

Classical set theory suggests that there are only two states of the properties of objects: belonging or not belonging, i.e., 1 or 0. Zadeh first proposed fuzzy set theory [43]. Fuzzy set theory measures the probability of a sample belonging to a category through the concept of fuzzy membership value. The fuzzy membership value varies between 0 and 1, with the closer to 0 indicating that a sample does not belong to that state, and the closer to 1 indicating that it is more likely to belong to that state [44]. By combining DLDA with the fuzzy set theory, a fuzzy direct linear discriminant analysis (FDLDA) algorithm is proposed. FDLDA will have an advantage over DLDA when dealing with overlapped data.

The operation steps of FDLDA are as follows (input: data matrix X; output: transformation matrix W_{FDLDA}):

- Step 1. Calculate matrix S_{fw} and S_{fb};
- Step 2. Apply the eigendecomposition on S_{fb} to get Y_f and D_{fb}, then $Z_f \leftarrow Y_f D_{fb}^{-1/2}$;
- Step 3. Compute the eigenvalues and eigenvectors of ${Z_f}^T S_{fw} Z_f$ to get D_{fw} and U_f;
- Step 4. Obtain the transformation matrix: $W_{FDLDA} = D_{fw}^{-1/2} {U_f}^T {Z_f}^T$.

In step 1, the within-class scattering matrix S_{fw} and the between-class scattering matrix S_{fb} are calculated as follows:

$$S_{fw} = \sum_{j=1}^{c} \sum_{i=1}^{n} U_{ij}^{m} (x_i - \overline{x}_j)(x_i - \overline{x}_j)^T \tag{4}$$

$$S_{fb} = \sum_{j=1}^{c} \sum_{i=1}^{n} U_{ij}^{m} (\overline{x}_j - \overline{x})(x_j - \overline{x})^T \tag{5}$$

Here, c represents the number of categories; n represents the number of training samples; U_{ij} represents the degree to which the ith sample data x_i belongs to category j, and the value of U_{ij} varies from 0 to 1; $\overline{x}$ represents the average value of the training samples; and $\overline{x}_j$ represents the average value of the jth category.

4. Classifiers

4.1. K-Nearest Neighbor (KNN) Classifier

KNN is a type of supervised classifier. It does not have a learning process in the general sense but rather uses the training data to partition the feature vector space, and the result of the partitioning is used as the final algorithmic model [45]. In order to categorize test samples, it first scans the training set in search of a training sample that is the most similar to the test sample. It then votes based on the class of this sample to establish the kind of test samples [46]. The classification accuracy of KNN is influenced by the parameters of K, distance metric, and the number of samples. To avoid identical votes in the poll, the parameter K is always set to an odd integer in the algorithm [47].

4.2. Extreme Learning Machine (ELM)

The extreme learning machine (ELM) stands as a prevalent classifier in the realm of pattern recognition [48]. The ELM can randomly initialize the input weights and biases, and once the weights and biases have been randomly determined, the output matrix of the hidden layer is uniquely determined [49]. This distinctive training process allows for the efficient learning of complex non-linear mapping relationships, endowing ELM with speed and versatility in model training [48].

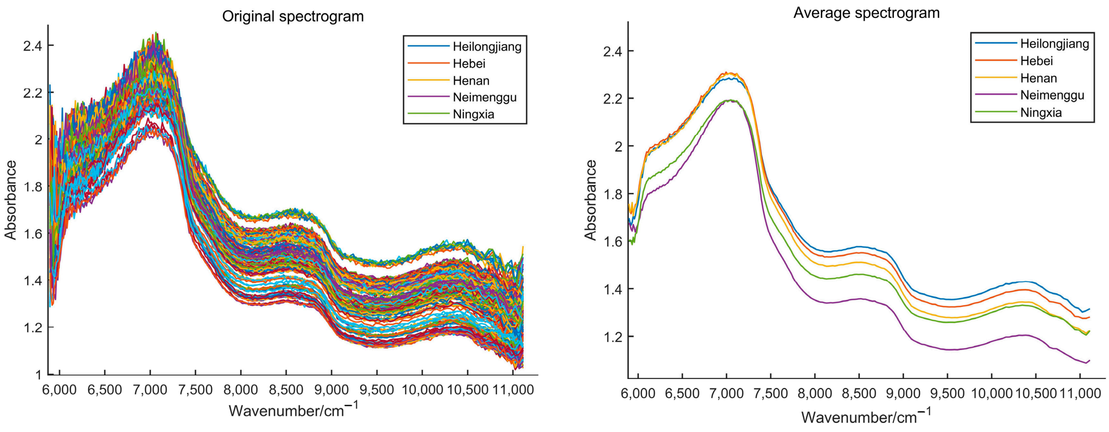
4.3. Naïve Bayes Classifier

The naïve Bayes classifier is a statistical classification algorithm based on Bayes' theorem. Throughout the training phase, the algorithm constructs the model through the estimation of both prior probabilities and conditional probabilities [50]. In the classification stage, it computes the posterior probability for each category and subsequently designates the category possessing the highest posterior probability as the classification outcome for the given sample.

5. Results and Discussion

5.1. Spectral Analysis

In this study, the NIR spectra from five different types of milk were obtained. Figure 1 shows the spectral data of milk samples from five different provinces. There is an absorption peak in the NIR spectral data of milk at 7112 cm^{-1}. There is an overlap in the spectra of milk from Neimenggu and Ningxia at 7112 cm^{-1}, as well as an overlap in the spectra of milk from Heilongjiang, Hebei, and Henan. The absorption peaks here are shifted by the absorption bands due to the difference in lactose content [51]. It can be inferred that there was a difference in lactose content between milk from Neimenggu and Ningxia, and milk from Heilongjiang, Hebei, and Henan. The fluctuation at 10,417 cm^{-1} is related to the O-H bond content in water [52]. It can be inferred that the content of water in milk from Ningxia and Henan is similar, while the content of water in milk from the other three origins is more different. The fluctuations at 6579 cm^{-1}, 6849 cm^{-1}, and 8833 cm^{-1} are affected by the N-H bond in the protein [53]. It can be inferred that the protein content of milk from different origins varies greatly. The first-order frequency doubling of the stretching vibration of the C-H bond in fat is related to the fluctuation at 6250 cm^{-1} [54]. It can be inferred that the fat content of milk from Heilongjiang, Hebei, and Henan is similar and very different from that of milk from Neimenggu and Ningxia. Meanwhile, other components in milk may also contain C-H, N-H, and O-H bonds, which can also lead to fluctuations in the spectral curves. So, there are differences in the components and content of milk from the five origins. In general, there are differences in the type and number of chemical bonds in commercial milk from five different origins. Therefore, we can distinguish milk from different origins based on fluctuations in the spectral curves.

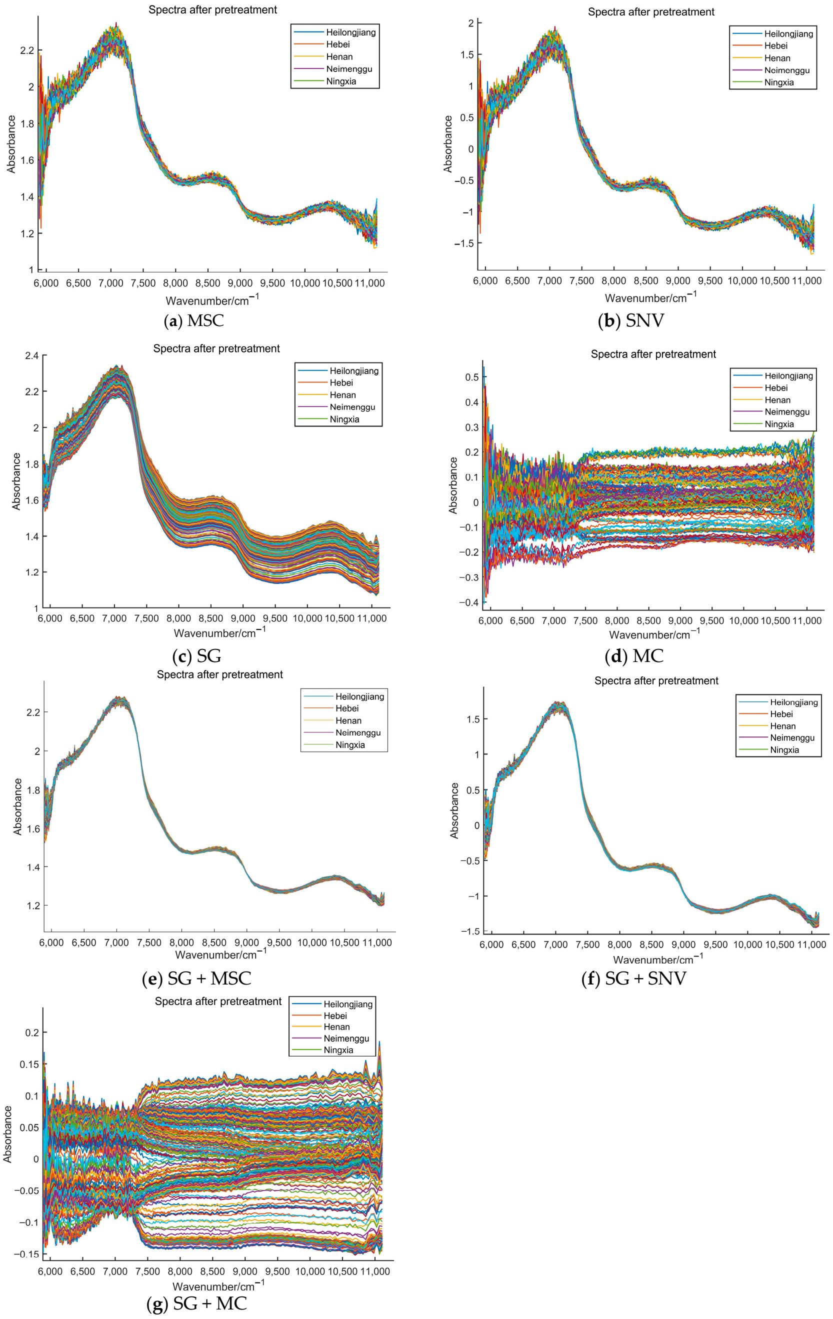


(**a**) Original spectrogram (**b**) Average spectrogram

Figure 1. The spectra of milk samples from 5 different provinces (**a**) original spectrogram; (**b**) average spec-trogram.

5.2. Results of Different Spectral Preprocessing

Figure 2 shows the NIR spectra of milk after some pretreatment procedures. MSC, SNV, SG, MC, SG + MSC, SG + SNV, and SG + MC were the data preprocessing methods used in this study. A comparison of (a), (b), (c), and (d) in Figure 2 showed that the spectrogram acquired by the SG pretreatment approach was smoother than those obtained by the other approaches. To see if better results could be obtained, the acquired data were processed using SG in conjunction with other preprocessing methods (e.g., SG + MSC, SG + SNV, and SG + MC). The spectrograms preprocessed by SG + MSC, SG + SNV, and SG + MC are shown in (e), (f), and (g).



(**a**) MSC (**b**) SNV

(**c**) SG (**d**) MC

(**e**) SG + MSC (**f**) SG + SNV

(**g**) SG + MC

Figure 2. NIR spectra of milk samples under different pretreatment methods: (**a**) multiplying scattering correction (MSC); (**b**) standard normal variables (SNV); (**c**) Savitzky–Golay (SG); (**d**) mean centering (MC); (**e**) Savitzky–Golay and multiplying scattering correction (SG + MSC); (**f**) Savitzky–Golay and standard normal variables (SG + SNV); (**g**) Savitzky–Golay and mean centering (SG + MC).

Table 4 demonstrates the classification accuracy of the model after using different preprocessing methods. It could be seen that the model's classification accuracies were low when the preprocessing methods were MSC, SNV, and MC, indicating that the spectra by the preprocessing methods in Figure 2a,b,d with high noise and distortion. The classification accuracies under the SG, SG + MSC, SG + SNV, and SG + MC preprocessing methods were high, and they were corroborated by the comparative smoothing of the spectra in Figure 2c,e,f,g with no significant noise. Table 4 also indicated that the classification accuracies of the combined two preprocessing methods were higher than those of only one preprocessing method.

Table 4. Classification accuracy of models with different preprocessing algorithms (%).

Feature Extraction	MSC	SNV	SG	MC	SG + MSC	SG + SNV	SG + MC
LDA	30.67	32.00	94.67	64.00	92.00	90.67	82.67
DLDA	38.67	40.00	90.67	60.00	93.00	94.67	92.00
FDLDA	34.67	34.67	94.67	66.67	96.00	97.33	94.67

Abbreviation: MSC: multiplying scattering correction; SNV: standard normal variables; SG: Savitzky–Golay; MC: mean centering; LDA: linear discriminant analysis; DLDA: direct linear discriminant analysis; FDLDA: fuzzy direct linear discriminant analysis.

5.3. PCA for Dimensionality Reduction

Principal component analysis (PCA) was employed to reduce the dimensionality of the NIR spectral data after the preprocessing of the milk spectral data. The number of major components was often determined by the cumulative contribution rate as a criterion. Table 5 illustrates the classification accuracies of LDA, DLDA, and FDLDA with different numbers of principal components. LDA had the maximum classification accuracy with 94.67% when the number of principal components (NPC) was set as seven. When the NPC was set as five, DLDA and FDLDA had the highest classification accuracies with 94.67% and 97.33%, respectively. However, when the NPC was set as 11, LDA and DLDA had relatively low classification accuracies. As a result, the NPC was set as five in this study.

Table 5. Classification accuracy under different numbers of principal components (%).

PCs	4	5	6	7	8	9	10	11	12
LDA	88.00	90.67	90.67	94.67	94.67	92.00	90.67	80.00	76.00
DLDA	94.67	94.67	93.33	92.00	94.67	89.33	86.67	89.33	82.67
FDLDA	94.67	97.33	89.33	93.33	90.67	94.67	94.67	97.33	92.00

Abbreviation: PCs: principal components; LDA: linear discriminant analysis; DLDA: direct linear discriminant analysis; FDLDA: fuzzy direct linear discriminant analysis.

When there were five principal components in this study, the total contribution of the first five principal components reached 99.99%, and they could roughly capture the feature information of all the data. The eigenvalues were as follows: $\lambda_1 = 66.3279$; $\lambda_2 = 15.2091$; $\lambda_3 = 3.5933$; $\lambda_4 = 2.8142$; $\lambda_5 = 1.8735$; $\lambda_6 = 1.1185$; $\lambda_7 = 0.8914$; $\lambda_8 = 0.7189$; $\lambda_9 = 0.6504$; $\lambda_{10} = 0.5339$; $\lambda_{11} = 0.3589$.

By dimensionality reduction of the spectral data through PCA, the data point distribution is shown in Figure 3. It could be seen that there was a large overlap of milk sample data from Heilongjiang, Henan, and Hebei, which increased the difficulty of the subsequent classification task.

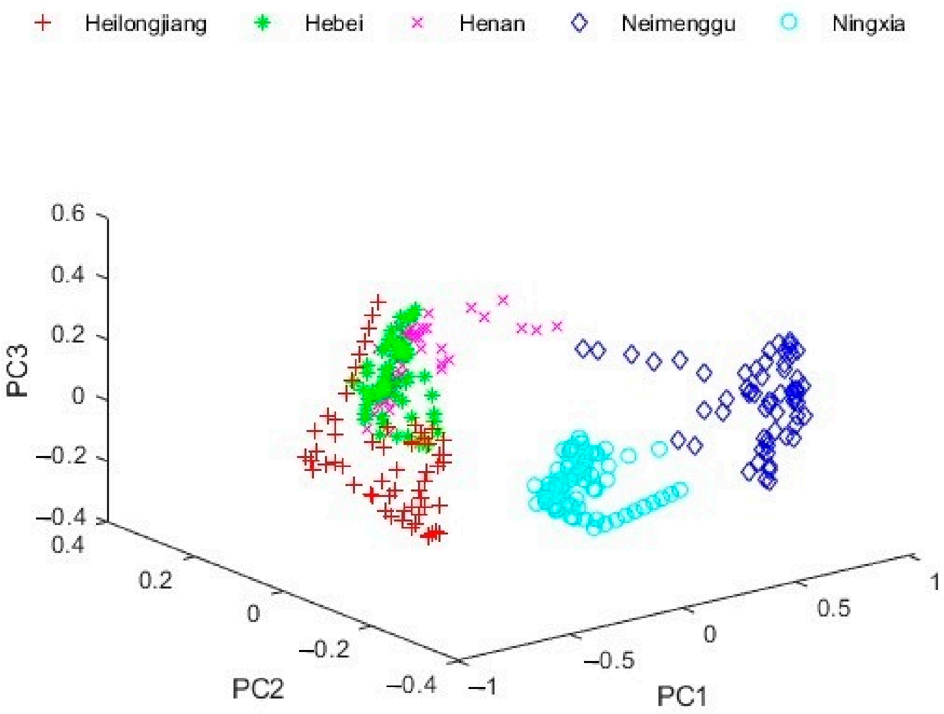


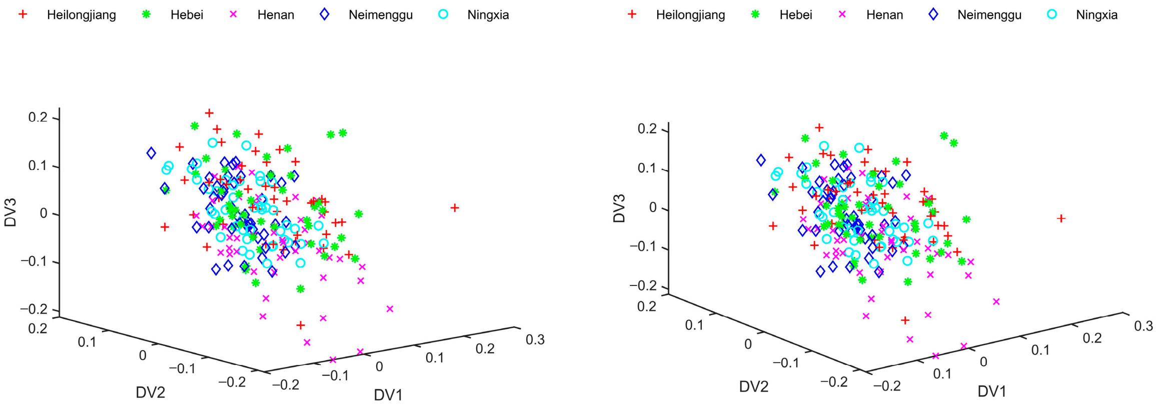
Figure 3. The data distribution after PCA.

5.4. Establishment of Discriminant Model

5.4.1. Feature Extraction with DLDA

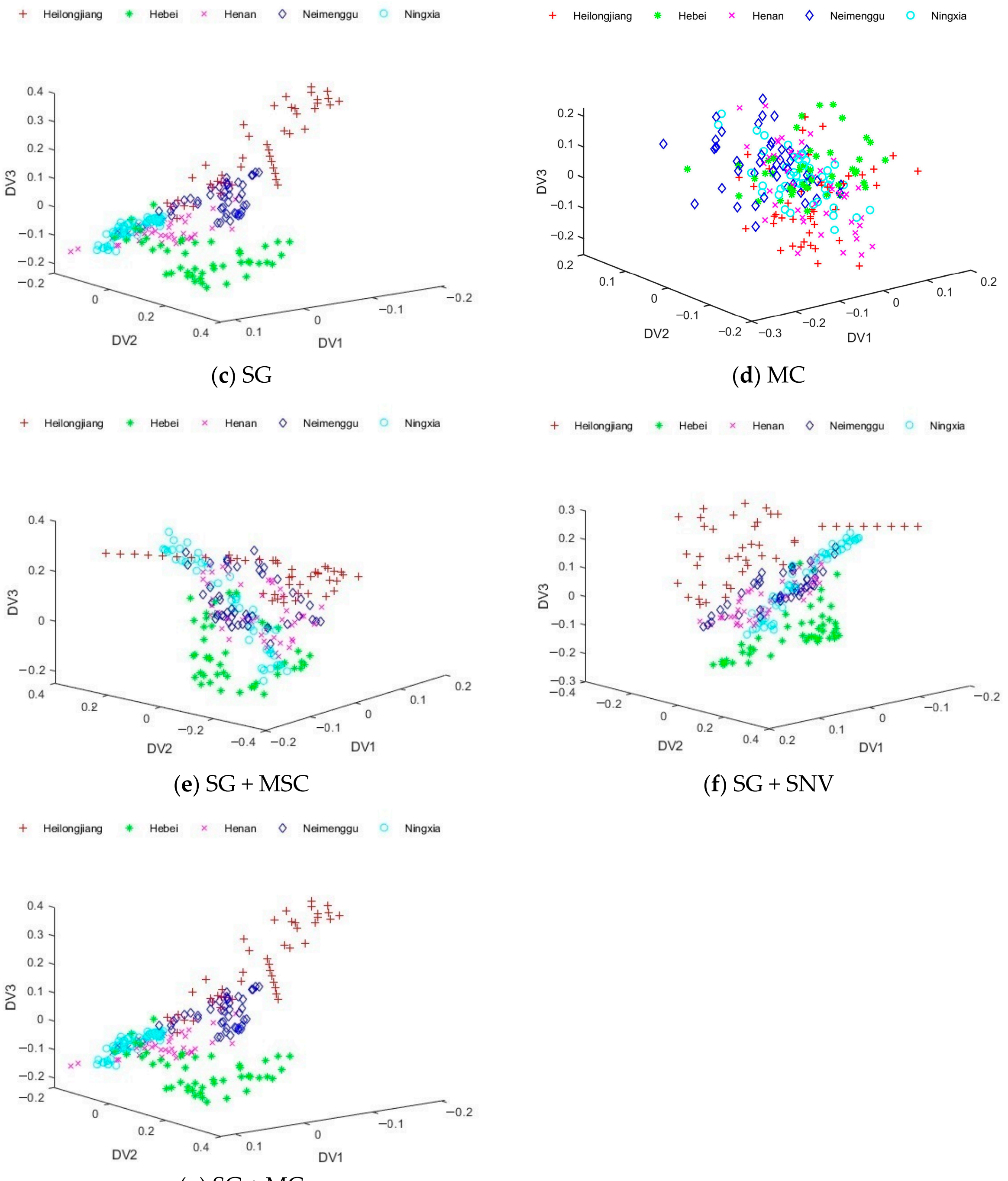
Since the PCA dimensionality reduction process set the NPCs to five, this meant that PCA reduced the original 228-dimensional data to five dimensions. For feature extraction using DLDA, the transformation matrix further reduced the dimensionality of the sample data to four dimensions (number of categories − 1). This process generated the four best discriminant vectors (DV1, DV2, DV3, and DV4).

As can be seen from Figure 4, the distribution of sample data after DLDA feature extraction was chaotic. Most of the milk samples from Heilongjiang and Hebei were able to be separated when pretreated with SG, SG + SNV, and SG + MC. However, there was a large overlap in milk samples from Henan, Neimenggu, and Ningxia. DLDA did not distinguish milk samples from different origins very well. In order to deal with these overlapped data, this study needed to introduce the fuzzy theory and develop a new feature extraction algorithm.



(**a**) MSC (**b**) SNV

Figure 4. *Cont.*



(**c**) SG

(**d**) MC

(**e**) SG + MSC

(**f**) SG + SNV

(**g**) SG + MC

Figure 4. The distribution of training samples after direct linear discriminant analysis (DLDA) under different preprocessing methods: (**a**) multiplying scattering correction (MSC); (**b**) standard normal variables (SNV); (**c**) Savitzky–Golay (SG); (**d**) mean centering (MC); (**e**) Savitzky–Golay and multiplying scattering correction (SG + MSC); (**f**) Savitzky–Golay and standard normal variables (SG + SNV); (**g**) Savitzky–Golay and mean centering (SG + MC). (There are 45 training samples for each type of commercial milk, totaling 225 training samples for the 5 commercial kinds of milk.).

5.4.2. Feature Extraction with FDLDA

As shown in Figure 5. The horizontal and vertical axes indicated the number of samples and the fuzzy membership value, respectively. The fuzzy membership value in Figure 5 was calculated using Euclidean distance, which indicated the probability that each of the 225 samples in the training set belonged to each category. When the fuzzy membership value was closer to 1, it meant that the probability of the sample belonging to that category was higher. Similarly, if the fuzzy membership value was closer to 0, it meant that the probability of the sample belonging to that category was less. For example, the fourth subplot of Figure 5 represented the probability of being from Ningxia among the 225 milk training samples. It could be seen that the fuzzy membership value of the 136th–180th samples was very close to 1, while the rest of the samples had a fuzzy membership value of basically 0. This indicated that the 136th–180th milk samples were more likely to be from Ningxia, while the rest of the samples were more likely to be from the other four provinces.

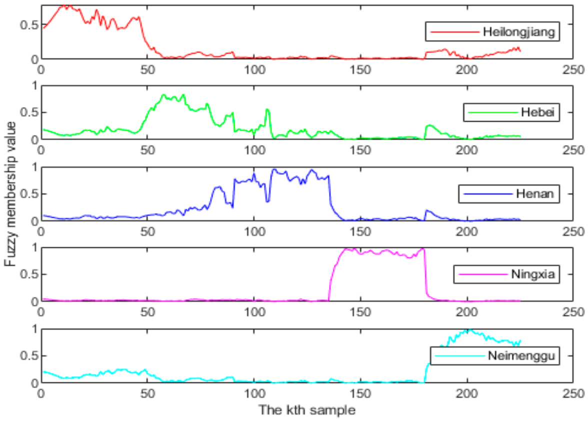
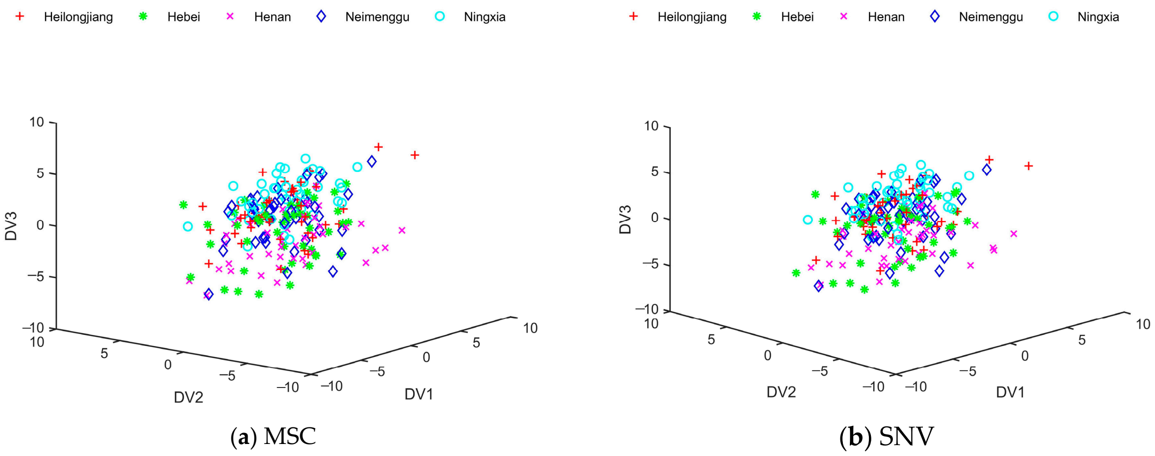


Figure 5. Fuzzy membership values of milk samples.

Figure 6 shows the distribution of results for extracting discrimination information from milk samples using FDLDA as a feature extraction technique. The figure showed a good separation of data points for the five types of samples. The study also found that the accuracies were improved by combining two preprocessing methods, such as SG + MSC, SG + SNV, and SG + MC, compared to using the SG algorithm alone. As shown in Figures 4 and 6, compared with DLDA, the distribution boundaries between different types of milk data processed by FDLDA were much clearer, which considerably solved the problem of data overlapping. The classification accuracy of FDLDA achieved 97.33% in the classification of milk samples from five distinct geographical origins.

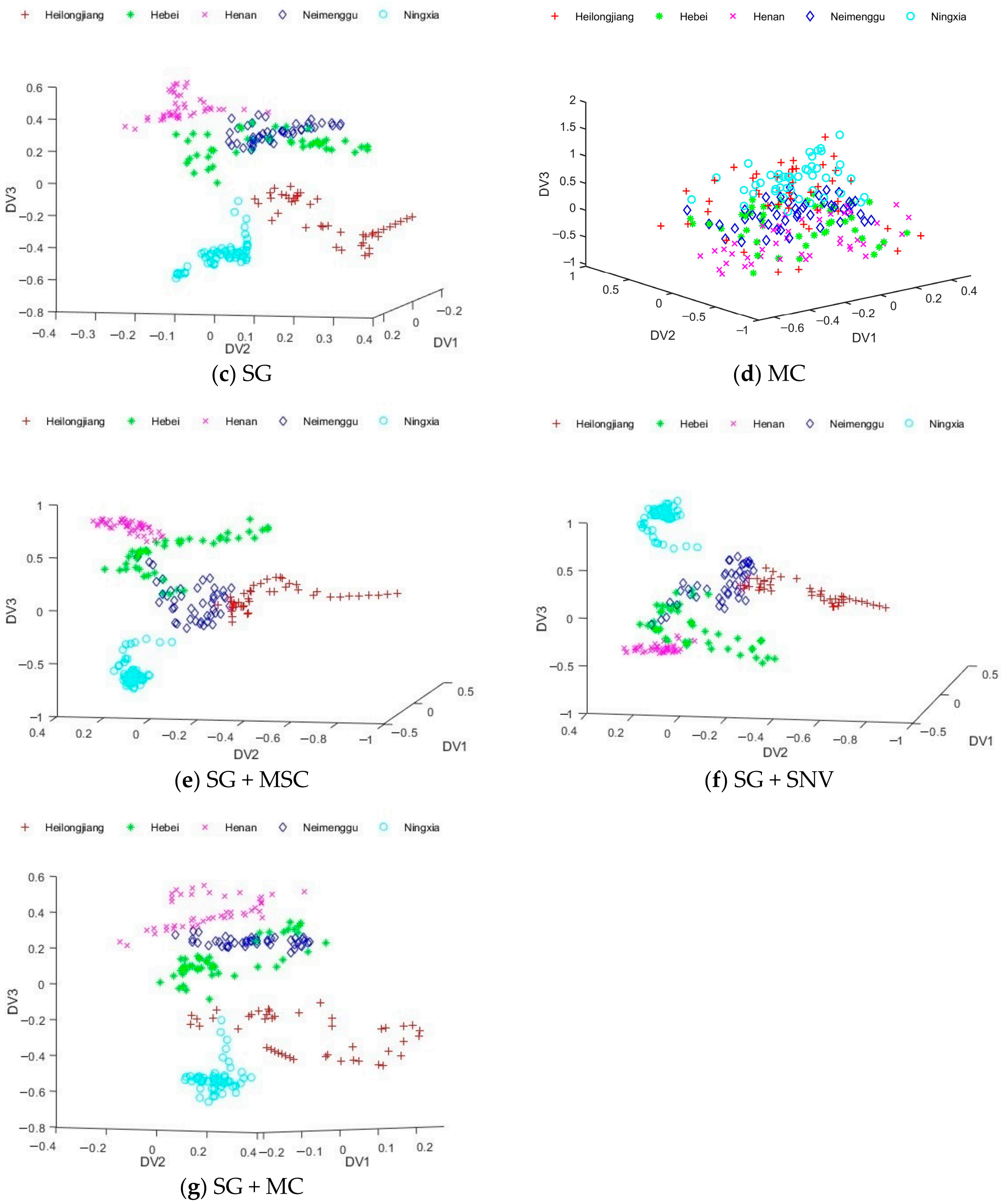
Combined with Table 4, it could also be seen that the classification accuracies of the FDLDA model were higher than those of the DLDA model when using SG, SG + SNV, SG + MSC, and SG + MC as preprocessing methods. In particular, it could be inferred that the ability of FDLDA to separate overlapped data was better than that of DLDA. Therefore, combining FDLDA with NIR spectroscopy could effectively determine the geographical origin of milk.



(**a**) MSC

(**b**) SNV

Figure 6. *Cont.*



(**c**) SG

(**d**) MC

(**e**) SG + MSC

(**f**) SG + SNV

(**g**) SG + MC

Figure 6. The distribution of training samples after fuzzy direct linear discriminant analysis (FDLDA) under different preprocessing methods: (**a**) multiplying scattering correction (MSC); (**b**) standard normal variables (SNV); (**c**) Savitzky–Golay (SG); (**d**) mean centering (MC); (**e**) Savitzky–Golay and multiplying scattering correction (SG + MSC); (**f**) Savitzky–Golay and standard normal variables (SG + SNV); (**g**) Savitzky–Golay and mean centering (SG + MC). (There are 45 training samples for each type of commercial milk, totaling 225 training samples for the 5 commercial kinds of milk.).

The fuzzy weight coefficient m had a significant impact on the final results when FDLDA was applied for feature extraction (SG + SNV was the preprocessing algorithm, and the classifier was KNN with parameter $K = 7$). It was difficult for FDLDA to separate the data if m was too small, and the fuzzy membership values of different kinds of data would become ambiguous if m was too large. Figure 7 shows the FDLDA model's classification accuracies for different fuzzy weight coefficient m values. The classification accuracy increased gradually as m increased. When the fuzzy weight coefficient m was 3.5, the classification accuracy reached its peak at 97.33%.

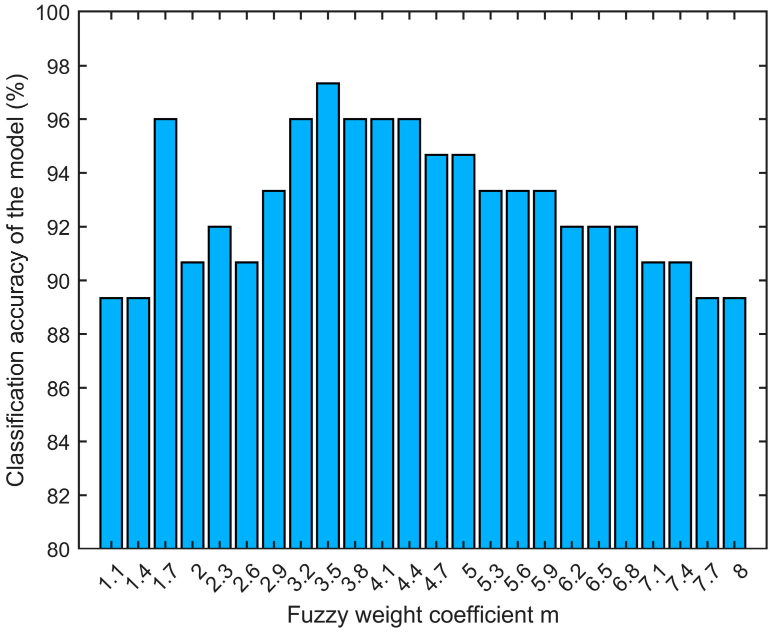


Figure 7. Classification results of FDLDA with different values of weight coefficient m.

5.5. Classification Results

5.5.1. Classification by KNN Classifier

This research integrated several data preprocessing methods and feature extraction algorithms to analyze the spectral data of milk samples and then applied a KNN classifier to achieve effective discrimination of the geographical source of milk. The choice of K value had a significant impact on the performance. Table 6 demonstrates the classification accuracies of LDA, DLDA, and FDLDA with different K values (the preprocessing method was SG + SNV; the number of principal components was 5; the fuzzy weight coefficient $m = 3.5$). As seen in Table 6, with the same K value, the classification accuracy of FDLDA was higher than LDA and DLDA. Furthermore, FDLDA achieved the highest recognition accuracy 97.33% when K = 7. The results presented above suggest that FDLDA was a more suitable algorithm for feature extraction when processing NIR spectra of milk samples, particularly in cases where the data exhibits some degree of overlap.

Table 6. Classification accuracy of the model using KNN with different K values (%).

K	**1**	**3**	**5**	**7**	**9**	**11**
LDA	85.33	88.00	93.33	90.67	88.00	90.67
DLDA	82.67	86.67	93.33	94.67	89.33	89.33
FDLDA	88.00	93.33	96.00	97.33	93.33	96.00

Abbreviation: LDA: linear discriminant analysis; DLDA: direct linear discriminant analysis; FDLDA: fuzzy direct linear discriminant analysis.

5.5.2. Classification by ELM

Subsequently, the classification performance of the FDLDA model was validated using an extreme learning machine (ELM). The number of hidden layer neurons was a crucial hyperparameter that directly impacted the model's performance and generalization ability. In this study, the optimal number of hidden layer neurons was selected from the range of 1 to 20. The optimal numbers of hidden layer neurons for the LDA, DLDA, and FDLDA models were determined to be 7, 16, and 10, respectively. Due to the stochastic nature of

ELM results, the average of 50 experimental results was taken as the final result. Table 7 displays the classification accuracies of the models under different preprocessing methods (the value of the fuzzy weight coefficient *m* was 3). The findings indicated that when employing SG, SG + MC, and SG + SNV as preprocessing methods, the LDA, DLDA, and FDLDA models achieved the highest classification accuracies. Notably, the FDLDA model exhibited superior classification accuracy compared to both the LDA and DLDA models.

Table 7. Classification accuracy of the model obtained using ELM with different preprocessing methods (%).

Feature Extraction	MSC	SNV	SG	MC	SG + MSC	SG + SNV	SG + MC
LDA	53.25	52.53	85.92	65.97	83.93	82.56	84.90
DLDA	52.40	51.70	84.96	66.88	81.01	81.04	86.03
FDLDA	51.45	51.87	87.14	66.67	87.25	88.24	86.42

Abbreviation: MSC: multiplying scattering correction; SNV: standard normal variables; SG: Savitzky–Golay; MC: mean centering; LDA: linear discriminant analysis; DLDA: direct linear discriminant analysis; FDLDA: fuzzy direct linear discriminant analysis; ELM, extreme learning machine.

5.5.3. Classification by Naïve Bayes Classifier

In the end, the classification performance of the FDLDA model was assessed by employing a naïve Bayes classifier. Table 8 displays the classification accuracy using the naïve Bayes classifier under different preprocessing methods, with the weight coefficient *m* set to 2. The highest classification accuracies of LDA, DLDA, and FDLDA models were more than 90%, specifically, 93.33%, 90.67%, and 92.00%, respectively. It was noteworthy that the FDLDA model exhibited higher classification accuracy compared to the DLDA model, indicating that FDLDA was better suited for handling overlapped data than DLDA.

Table 8. Classification accuracy of the model using naïve Bayes classifier with different preprocessing methods (%).

Feature Extraction	MSC	SNV	SG	MC	SG + MSC	SG + SNV	SG + MC
LDA	60.00	60.00	88.00	70.67	93.33	93.33	88.00
DLDA	46.67	46.67	90.67	61.33	90.67	90.67	90.67
FDLDA	52.00	52.00	89.33	66.67	92.00	92.00	89.33

Abbreviation: MSC: multiplying scattering correction; SNV: standard normal variables; SG: Savitzky–Golay; MC: mean centering; LDA: linear discriminant analysis; DLDA: direct linear discriminant analysis; FDLDA: fuzzy direct linear discriminant analysis.

5.6. Comparative Analysis

The model FDLDA-KNN proposed in this study achieved a classification accuracy of 97.33% for milk test samples. Marijan et al. successfully differentiated between sheep's milk from various geographical origins by accurately determining the stable isotope ratios and elemental composition of the primary bio-elements, achieving an impressive accuracy rate of 97.00% [55]. Liu et al. successfully classified goat milk from three pastures in China. This accomplishment was realized through the precise determination of lipid molecules present in goat milk samples from these regions, resulting in an impeccable classification accuracy of 100% [56]. Ng et al. determined the stable isotope signature and elemental concentration of milk to differentiate between milk from Australia, New Zealand, Thailand, and the USA, achieving a classification accuracy of 94.7% [12]. Magdas et al. studied the isotopic and elemental composition of raw milk from Transylvania and successfully differentiated raw milk from three pastures with 90.9% classification accuracy [57].

The comparison shows that most of the origin traceability for milk was performed by determining its composition. Although achieving high accuracy, this method is complex, time-consuming, and lossy, making it difficult to commercialize on a large scale. In contrast, the origin traceability model proposed in this study is fast and non-destructive while still achieving high classification accuracy. This provides a reference for future research.

6. Conclusions

In this study, the NIR spectral data of the milk were acquired by a portable NIR spectrometer. The data were preprocessed by different kinds of data preprocessing methods, and then PCA was applied to reduce the dimensionality of the data. LDA, DLDA, and FDLDA were employed to obtain feature information from the NIR spectra of milk. Finally, the KNN classifier, ELM, and naïve Bayes classifier were applied to classify five types of milk samples. Experimental results indicated that the KNN classifier outperformed ELM and the naïve Bayes classifier in terms of classification accuracy. FDLDA-KNN achieved the highest classification accuracy of 97.33% for the milk test samples. This meant that the FDLDA-KNN model showed greater ability than the DLDA-KNN model to handle overlapped data, clarify data boundaries, and achieve the highest classification accuracy. In summary, the combination of FDLDA with near-infrared spectroscopy offered an effective method for identifying the geographical origin of milk.

Milk from different origins varies in terms of its composition and content, which results in different types and amounts of chemical bonds. NIR spectroscopy can detect the expansion and contraction vibration of different chemical bonds, so as long as the milks are from different origins, there will be differences in their spectra. And the algorithm proposed in this study first learns using the training data to extract discriminant information and then classifies the test data. So, this study can be generalized to a large extent to milk from other origins and it is not limited by the origins selected in this study.

At present, the integration of NIR spectroscopy technology with other technologies has become the mainstream of food origin traceability research. NIR spectroscopy has played an important role in the food industry, but the mechanism of origin traceability and the establishment of various discriminative models are still in the exploratory stage. A complete traceability system has not yet been established. This study provides a reference for the traceability of other dairy products or food in the future. The research on the origin traceability of meat products is less, and the geographical origin of each field needs to be studied from multiple angles in the future. It is also possible to think about the combination of spectroscopic techniques, chemometrics methods, and other techniques (e.g., radio frequency identification) to build a complete food traceability system.

Author Contributions: Conceptualization, B.W. and X.W.; methodology, X.W. and T.Z.; software, T.Z., B.W. and Y.W.; validation, B.W., Y.W. and X.W.; formal analysis, C.H.; investigation, C.H. and Y.W.; resources, B.W. and X.W.; data curation, J.S.; writing—original draft preparation, Y.W. and C.H.; writing—review and editing, X.W. and B.W.; visualization, J.S.; supervision, J.S. and X.W.; project administration, X.W. and B.W.; funding acquisition, B.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Natural Science Research Projects of Colleges and Universities in Anhui Province (2022AH040333), and the Youth and Middle-aged Teachers Cultivation Action Project in Anhui Province (JNFX2023136).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

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

References

1. Pereira, P.C. Milk nutritional composition and its role in human health. *Nutrition* **2014**, *30*, 619–627. [CrossRef] [PubMed]
2. Bemfeito, R.M.; Rodrigues, J.F.; Silva, J.G.E.; Abreu, L.R. Temporal dominance of sensations sensory profile and drivers of liking of artisanal Minas cheese produced in the region of Serra da Canastra, Brazil. *J. Dairy Sci.* **2016**, *99*, 7886–7897. [CrossRef] [PubMed]
3. Hebling, E.; Tavares, J.P.; Da Silva Medeiros, M.L.; Barbin, D.F. Near-infrared techniques for fraud detection in dairy products: A review. *J. Food Sci.* **2022**, *87*, 1943–1960. [CrossRef] [PubMed]

4. Maitiniyazi, S.; Canavari, M. Exploring Chinese consumers' attitudes toward traceable dairy products: A focus group study. *J. Dairy Sci.* **2020**, *103*, 11257–11267. [CrossRef] [PubMed]
5. Esteki, M.; Shahsavari, Z.; Simal-Gandara, J. Use of spectroscopic methods in combination with linear discriminant analysis for authentication of food products. *Food Control* **2018**, *91*, 100–112. [CrossRef]
6. Esteki, M.; Shahsavari, Z.; Simal-Gandara, J. Food identification by high performance liquid chromatography fingerprinting and mathematical processing. *Food Res. Int.* **2019**, *122*, 303–317. [CrossRef]
7. Katerinopoulou, K.; Kontogeorgos, A.; Salmas, C.E.; Patakas, A.; Ladavos, A. Geographical origin authentication of agri-food products: A review. *Foods* **2020**, *9*, 489. [CrossRef]
8. Wadood, S.A.; Boli, G.; Xiaowen, Z.; Hussain, I.; Yimin, W. Recent development in the application of analytical techniques for the traceability and authenticity of food of plant origin. *Microchem. J.* **2020**, *152*, 104295. [CrossRef]
9. Wu, J.; Zareef, M.; Chen, Q.; Ouyang, Q. Application of visible-near infrared spectroscopy in tandem with multivariate analysis for the rapid evaluation of matcha physicochemical indicators. *Food Chem.* **2022**, *421*, 136185. [CrossRef]
10. Zhang, X.; Wang, Y.; Zhou, Z.; Zhang, Y.; Wang, X. Detection method for tomato leaf mildew based on hyperspectral fusion terahertz technology. *Foods* **2023**, *12*, 535. [CrossRef]
11. Luo, D.; Dong, H.; Luo, H.; Xian, Y.; Wan, J.; Guo, X.; Wu, Y. The application of stable isotope ratio analysis to determine the geographical origin of wheat. *Food Chem.* **2015**, *174*, 197–201. [CrossRef] [PubMed]
12. Ng, W.L.; Bay, L.J.; Goh, G.; Ang, T.H.; Kong, K.; Chew, P.; Koh, S.P.; Ch'ng, A.L.; Phang, H.; Chiew, P. Multivariate statistical analysis of stable isotope signatures and element concentrations to differentiate the geographical origin of retail milk sold in Singapore. *Food Control* **2021**, *123*, 107736. [CrossRef]
13. Kharbach, M.; Viaene, J.; Yu, H.; Kamal, R.; Marmouzi, I.; Bouklouze, A.; Heyden, Y.V. Secondary-metabolites fingerprinting of Argania Spinosa kernels using liquid chromatography–mass spectrometry and chemometrics, for metabolite identification and quantification as Well as for geographic classification. *J. Chromatogr. A* **2022**, *1670*, 462972. [CrossRef] [PubMed]
14. Xie, L.; Zhao, S.; Rogers, K.M.; Xia, Y.; Zhang, B.; Suo, R.; Zhao, Y. A case of milk traceability in small-scale districts-Inner Mongolia of China by nutritional and geographical parameters. *Food Chem.* **2020**, *316*, 126332. [CrossRef] [PubMed]
15. Zhang, D.; Wu, W.; Qiu, X.; Li, X.; Zhao, F.; Ye, N. Rapid and direct identification of the origin of white tea with proton transfer reaction-time of flight-mass spectrometry. *Rapid Commun. Mass Spectrom.* **2020**, *34*, e8830. [CrossRef] [PubMed]
16. Zuo, Y.; Tan, G.; Xiang, D.; Chen, L.; Wang, J.; Zhang, S.; Bai, Z.; Wu, Q. Development of a novel green tea quality roadmap and the complex sensory-associated characteristics exploration using rapid near-infrared spectroscopy technology. *Spectrochim. Acta A* **2021**, *258*, 119847. [CrossRef] [PubMed]
17. Dong, C.; Liu, Z.; Yang, C.; An, T.; Hu, B.; Luo, X.; Jin, J.; Li, Y. Rapid detection of exogenous sucrose in black tea samples based on near-infrared spectroscopy. *Infrared Phys. Technol.* **2021**, *119*, 103934. [CrossRef]
18. Qi, Z.; Wu, X.; Yang, Y.; Wu, B.; Fu, H. Discrimination of the red jujube varieties using a portable NIR spectrometer and fuzzy improved linear discriminant analysis. *Foods* **2022**, *11*, 763. [CrossRef] [PubMed]
19. Luykx, D.M.A.M.; Van Ruth, S.M. An overview of analytical methods for determining the geographical origin of food products. *Food Chem.* **2008**, *107*, 897–911. [CrossRef]
20. Zhang, L.-G.; Zhang, X.; Ni, L.-J.; Xue, Z.-B.; Gu, X.; Huang, S.-X. Rapid identification of adulterated cow milk by non-linear pattern recognition methods based on near infrared spectroscopy. *Food Chem.* **2014**, *145*, 342–348. [CrossRef]
21. Diaz-Olivares, J.A.; Adriaens, I.; Stevens, E.; Saeys, W.; Aernouts, B. Online milk composition analysis with an on-farm near-infrared sensor. *Comput. Electron. Agric.* **2020**, *178*, 105734. [CrossRef]
22. Pereira, E.V.D.S.; Fernandes, D.D.D.S.; De Araújo, 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* **2020**, *127*, 109427. [CrossRef]
23. Coppa, M.; Martin, B.; Agabriel, C.; Chassaing, C.; Sibra, C.; Constant, I.; Graulet, B.; Andueza, D. Authentication of cow feeding and geographic origin on milk using visible and near-infrared spectroscopy. *J. Dairy Sci.* **2012**, *95*, 5544–5551. [CrossRef] [PubMed]
24. Yin, X. The Regression Analysis and Pattern Recognition of Milk Based on Infrared Spectroscopy. Master's Thesis, Jiangnan University, Wuxi, China, June 2013.
25. Guo, Z.; Jayan, H. Fast nondestructive detection technology and equipment for food quality and safety. *Foods* **2023**, *12*, 3744. [CrossRef] [PubMed]
26. Liu, T.; He, J.; Yao, W.; Jiang, H.; Chen, Q. Determination of aflatoxin B1 value in corn based on Fourier transform near-infrared spectroscopy: Comparison of optimization effect of characteristic wavelengths. *LWT* **2022**, *164*, 113657. [CrossRef]
27. Zhang, T.; Wu, X.; Wu, B.; Dai, C.; Fu, H. Rapid authentication of the geographical origin of milk using portable near-infrared spectrometer and fuzzy uncorrelated discriminant transformation. *J. Food Process Eng.* **2022**, *45*, e14040. [CrossRef]
28. Sharma, A.; Paliwal, K.K. Linear discriminant analysis for the small sample size problem: An overview. *Int. J. Mach. Learn.* **2015**, *6*, 443–454. [CrossRef]
29. Portillo-Portillo, J.; Leyva, R.; Sanchez, V.; Sanchez-Perez, G.; Perez-Meana, H.; Olivares-Mercado, J.; Toscano-Medina, K.; Nakano-Miyatake, M. A view-invariant gait recognition algorithm based on a joint-direct linear discriminant analysis. *Appl. Intell.* **2017**, *48*, 1200–1217. [CrossRef]
30. Paliwal, K.K.; Sharma, A. Improved direct LDA and its application to DNA microarray gene expression data. *Pattern Recognit. Lett.* **2010**, *31*, 2489–2492. [CrossRef]

31. Patel, H.; Thakur, G.S. An improved fuzzy k-nearest neighbor algorithm for imbalanced data using adaptive approach. *IETE J. Res.* **2019**, *65*, 780–789. [CrossRef]
32. Wang, J. Examination on face recognition method based on type 2 blurry. *JIFS* **2020**, *38*, 3929–3938. [CrossRef]
33. Shen, L.; Wang, H.; Luo, X.; Ni, J.; Ma, Y.; Chen, M.; Xiong, Q.; Zhang, S. Fatty acids and conventional nutrients in milk and their influencing factors. *China Dairy Cattle* **2021**, *2*, 45–51. [CrossRef]
34. Guo, H. Construction of Fingerprint of Specific Quality Characters of Milk in Ningxia of China. Master's Thesis, Ningxia University, Ningxia, China, April 2023. [CrossRef]
35. Wang, M.; Gao, M.; Li, Y.; Liu, J.; Fu, H.; Sun, X.; Duan, S. Traceability technology of milk origin based on stable isotope and mineral elements. *Sci. Technol. Food Ind.* **2022**, *3*, 284–290. [CrossRef]
36. Zhao, S.; Zhao, Y.; Rogers, K.M.; Chen, G.; Chen, A.; Yang, S. Application of Multi-Element (C, N, H, O) Stable Isotope Ratio Analysis for the Traceability of Milk Samples from China. *Food Chem.* **2020**, *310*, 125826. [CrossRef] [PubMed]
37. Xie, L. Study on Origin of Milk Based on the Characteristics of Stable Isotopes, Mineral Elements and Amino Acids. Master's Thesis, Hebei Agricultural University, Baoding, China, May 2020. [CrossRef]
38. He, F.; Wu, X.; Wu, B.; Zeng, S.; Zhu, X. Green tea grades identification via Fourier transform near-infrared spectroscopy and weighted global fuzzy uncorrelated discriminant transform. *J. Food Process Eng.* **2022**, *45*, e14109. [CrossRef]
39. Shen, Y.; Wu, X.; Wu, B.; Tan, Y.; Liu, J. Qualitative analysis of lambda-cyhalothrin on Chinese cabbage using mid-infrared spectroscopy combined with fuzzy feature extraction algorithms. *Agriculture* **2021**, *11*, 275. [CrossRef]
40. Rezghi, M.; Obulkasim, A. Noise-free principal component analysis: An efficient dimension reduction technique for high dimensional molecular data. *Expert Syst. Appl.* **2014**, *41*, 7797–7804. [CrossRef]
41. Howley, T.; Madden, M.G.; O'Connell, M.-L.; Ryder, A.G. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowl.-Based Syst.* **2006**, *19*, 363–370. [CrossRef]
42. Yu, H.; Yang, J. A direct LDA algorithm for high-dimensional data—With application to face recognition. *Pattern Recogni.* **2001**, *34*, 2067–2070. [CrossRef]
43. Zadeh, L.A. Similarity relations and fuzzy orderings. *Inf. Sci.* **1971**, *3*, 177–200. [CrossRef]
44. Bezdek, J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Springer: Boston, MA, USA, 1981.
45. Gómez-Meire, S.; Campos, C.; Falqué, E.; Díaz, F.; Fdez-Riverola, F. Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques. *Food Res. Int.* **2014**, *60*, 230–240. [CrossRef]
46. Ali, A.; Hamraz, M.; Kumam, P.; Khan, D.M.; Khalil, U.; Sulaiman, M.; Khan, Z. A k-nearest neighbours based ensemble via optimal model selection for regression. *IEEE Access* **2020**, *8*, 132095–132105. [CrossRef]
47. Nikith, B.V.; Keerthan, N.K.S.; Praneeth, M.S.; Amrita, T. Leaf disease detection and classification. *Procedia Comput.* **2023**, *218*, 291–300. [CrossRef]
48. Huang, G.-B.; Zhou, H.; Ding, X.; Zhang, R. Extreme learning machine for regression and multiclass classification. *IEEE Trans. Syst. Man Cybern. B* **2012**, *42*, 513–529. [CrossRef] [PubMed]
49. Ding, S.; Zhao, H.; Zhang, Y.; Xu, X.; Nie, R. Extreme learning machine: Algorithm, theory and applications. *Artif. Intell. Rev.* **2015**, *44*, 103–115. [CrossRef]
50. Menevseoglu, A.; Entrenas, J.A.; Gunes, N.; Dogan, M.A.; Pérez-Marín, D. Machine learning-assisted near-infrared spectroscopy for rapid discrimination of apricot kernels in ground almond. *Food Control* **2024**, *159*, 110272. [CrossRef]
51. Mohamed, H.; Nagy, P.; Agbaba, J.; Kamal-Eldin, A. Use of near and mid infra-red spectroscopy for analysis of protein, fat, lactose and total solids in raw cow and camel milk. *Food Chem.* **2021**, *334*, 127436. [CrossRef] [PubMed]
52. Ejeahalaka, K.K.; Mclaughlin, P.; On, S.L.W. Monitoring the composition, authenticity and quality dynamics of commercially available nigerian fat-filled milk powders under inclement conditions using NIRS, chemometrics, packaging and microbiological parameters. *Food Chem.* **2021**, *339*, 127844. [CrossRef] [PubMed]
53. Kang, R.; Wang, X.; Zhao, M.; Henihan, L.E.; O'Donnell, C.P. A comparison of benchtop and micro NIR spectrometers for infant milk formula powder storage time discrimination and particle size prediction using chemometrics and denoising methods. *J. Food Eng.* **2022**, *329*, 111087. [CrossRef]
54. Ashie, A.; Lei, H.; Han, B.; Xiong, M.; Yan, H. Fast determination of three components in milk thistle extract with a hand-held NIR spectrometer and chemometrics tools. *Infrared Phys. Tech.* **2021**, *113*, 103629. [CrossRef]
55. Nečemer, M.; Potočnik, D.; Ogrinc, N. Discrimination between Slovenian cow, goat and sheep milk and cheese according to geographical origin using a combination of elemental content and stable isotope data. *J. Food Compos. Anal.* **2016**, *52*, 16–23. [CrossRef]
56. Liu, H.; Guo, X.; Zhao, Q.; Qin, Y.; Zhang, J. Lipidomics analysis for identifying the geographical origin and lactation stage of goat milk. *Food Chem.* **2020**, *309*, 125765. [CrossRef] [PubMed]
57. Magdas, D.-A.; Dehelean, A.; Feher, I.; Cristea, G.; Puscas, R.; Dan, S.-D.; Cordea, D.-V. Discrimination markers for the geographical and species origin of raw milk within Romania. *Int. Dairy J.* **2016**, *61*, 135–141. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.