applied sciences

MDPI

Article

Portable Protein and Fat Detector in Milk Based on Multi-Spectral Sensor and Machine Learning

Yanyan Wang [1], Kaikai Zhang [1], Shengzhe Shi [1], Qingqing Wang [1] and Sheng Liu [1,2,*]

[1] School of Computer Science and Technology, Huaibei Normal University, Huaibei 235000, China; 12212080824@chnu.edu.cn (Y.W.); 12212080827@chnu.edu.cn (K.Z.); 12312080821@chnu.edu.cn (Q.W.)
[2] Engineering Research Center of Cognitive Behavioral Intelligent Computing and Application, Huaibei 235000, China
* Correspondence: liusheng@chnu.edu.cn

Abstract: To address the challenges of a long measurement period, high testing cost, and environmental pollution of traditional milk composition detection methods, a portable detection instrument was developed by combining multi-spectral sensors, machine learning algorithms, and an embedded system to rapidly detect the main components of milk. A broadband near-infrared (NIR) LED constant-current driver circuit and multi-spectral sensor module were designed to obtain six NIR features of milk samples. Based on a comparison of several machine learning algorithms, the XGBoost model was selected for training, and the trained model was ported to a Raspberry Pi unit for sample detection. The validation results showed that the coefficients of determination (R^2) for the investigated protein and fat models were 0.9816 and 0.9978, respectively, and the corresponding mean absolute errors (*MAE*) were 0.0086 and 0.0079. Accurate measurement of protein and fat contents of milk can be facilitated in a short time interval by using the proposed low-cost portable instrument.

Keywords: milk detection; machine learning; multi-spectral sensor; embedded system

Citation: Wang, Y.; Zhang, K.; Shi, S.; Wang, Q.; Liu, S. Portable Protein and Fat Detector in Milk Based on Multi-Spectral Sensor and Machine Learning. *Appl. Sci.* **2023**, *13*, 12320. https://doi.org/10.3390/app132212320

Academic Editor: Agata Górska

Received: 11 October 2023
Revised: 11 November 2023
Accepted: 13 November 2023
Published: 14 November 2023

1. Introduction

Milk is an important dairy product that is widely used in the food and nutrition industry, and accurately determining its contents is important in ensuring food quality and safety. Protein and fat are important components of milk, and protein plays a crucial role in the growth and development of the human body as well as the immune system [1,2]. Fat is a high-energy nutrient [3]. Therefore, in the process of milk procurement and production management, it is necessary to detect the contents of the main components of milk accurately and quickly, as this can not only provide a reference basis for milk quality analysis and production process quality control, but also provide scientific guidance for the excellent rearing of cows [4].

At present, classic detection methods are usually based on chemical analysis with typical standard detection methods such as the Kjeldahl nitrogen determination and Gable methods. The Kjeldahl method [5] calculates protein content by measuring the total nitrogen content in milk, whereas the Gable method [6] separates fat by adding sulfuric acid to calculate fat content. By contrast, traditional chemical analysis methods require longer detection time and complex chemical reactions, consume a large amount of reagents, and can also damage milk samples. This not only causes chemical pollution to the environment but also cannot meet the needs of online rapid measurement. Instrument analysis technologies, such as spectral technology, are widely used in milk composition analysis [7].

Optical testing methods are commonly used in non-destructive testing, mainly using infrared light, ultraviolet light, and fluorescence. Middle infrared (MIR) spectroscopy and near-infrared (NIR) spectroscopy are used extensively in the analysis of milk components [8,9]. Soyeurt et al. [10] used the MIR method to measure the fatty acid content in milk. Bonfatti et al. [11] employed MIR data from milk to predict the protein and fat content in

milk using Bayesian regression. Dabrowska et al. [12] applied broadband laser mid-infrared spectroscopy with quantum cascade detectors for milk protein analysis. These studies provide strong support for the application of MIR in the field of milk analysis. Although MIR technology has significant advantages in providing more chemical information and high resolution [13], it requires relatively complex instruments and operations and requires high transparency of samples, which may have certain limitations in practical applications. By contrast, near-infrared (NIR) spectroscopy technology is more suitable for the milk and dairy industry [14,15].

NIR has the characteristics of fast analysis speed and easy online real-time monitoring [16], meeting the needs of on-site measurement [17]. It is efficient, fast, and non-destructive [18], thus having great potential in the development of spectroscopic instruments [19]. Coppa et al. [20] predicted the composition of milk fatty acids using NIR technology. Diaz Olivares et al. [21] designed and established an online milk composition analysis system using NIR technology, with determination coefficients R^2 of 0.947 and 0.989 for protein and fat prediction models, respectively. Saranwong et al. [22] developed a system for the quality and safety assessment of heterogeneous raw milk using NIR technology. Mohamed et al. [9] employed NIR spectroscopy technology to analyze the protein and fat content in milk, with a wavelength range of 600–1050 nm and standard error values for fat and protein prediction of 0.25% and 0.15%, respectively. These results demonstrate the excellent performance of the NIR method in milk quality assessment.

Most NIR spectrometers are expensive and not suitable for a wide range of daily applications. In recent years, significant progress has been made in spectroscopic instruments, achieving low-cost miniaturized infrared instruments [21]. Muñoz-Salinas et al. [23] used a portable fluorescence detector to determine the protein content in milk. Zaky et al. [24] proposed a novel biophoton sensor based on porous silicon ternary photonic crystals for more effective detection of fat concentration in milk. Alamwgan et al. [25] used surface plasmon resonance structural optics based on MXene to detect fat concentration in milk, which was the first study to apply surface plasmon resonance to detect the fat concentration in milk. Yang et al. [26] developed a portable milk composition detector based on an NIR micro-spectrometer, which has driven the development of small spectral instrument equipment. These studies further indicate that although spectral instruments still have a certain cost and complexity, designing a portable milk composition detection instrument with fast detection speed and easy on-site use is of great significance in milk quality control and production. With the continuous development of micro-spectral instruments, it has become possible to achieve detection speeds comparable to laboratory instruments.

The aim of this study Is to design a portable instrument that can quickly detect the protein and fat content in milk using new sensor technology, machine learning algorithms, and embedded systems. We designed a broadband near-infrared LED light source driver circuit and a multi-wavelength sensor module, combined with the XGBoost machine learning model, to achieve highly accurate detection, overcoming the problems of traditional methods such as a long measurement cycle, high cost, and high environmental pollution. The development of a portable milk protein and fat detector is of great significance as it provides an efficient, fast, green, and low-cost method for quickly measuring the protein and fat content in milk on-site. This not only helps to improve milk quality control but also provides practical tools for farms, dairy factories, and food testing departments, helping to improve the production and processing processes of the milk industry. In addition, this technology can also be applied to other fields, providing convenient solutions for food analysis and quality control, and has broad application prospects, which can promote the development of portable instruments for milk quality testing.

2. Materials and Methods

2.1. Milk Samples

Sixty different types of milk samples were purchased from local supermarkets to construct a dataset (see Supplementary Materials Table S1 for more information on the milk samples) and stored in a refrigerator at 15 °C. Twenty sets of light-intensity data were collected from six channels (610, 680, 730, 760, 810, and 860 nm) for each milk sample as input labels for the model. The protein and fat contents were measured according to international standards and were used as output labels to construct an XGBoost milk composition prediction model. Each sample had six light-intensity features and two output features. The milk sample dataset used in the experiment contained 60 × 20 data points. First, 58 samples were randomly selected from the 60 milk samples, and 19 datasets were selected from each sample for a total of 58 × 19 data points as training set 1. The remaining 58 × 1 datasets were used as Test Set 1. The untrained 2 milk samples with 2 × 20 datasets were used as Test Set 2. Training set 1 was used to verify the accuracy of the model, and Test Set 2 was used to verify its generalization ability.

2.2. Measuring Device

Figure 1 shows the overall architecture of the portable instrument for milk detection based on a multi-spectral sensor. The instrument consists primarily of hardware and software. The hardware includes a Raspberry Pi (Raspberry Pi 4 B) (Shenzhen, China) development board, touch screen, lithium power module, cuvette, multi-spectral sensor, and broadband NIR LED controlled by a constant-current driver circuit. PyQt5 (Version 5.15) and Python codes (Version 3.7) were used to develop the software interface and implement the required functions to train the model to detect the main components of milk. Integration of the hardware and software components resulted in a portable instrument for milk composition measurements.

Multispectral sensor (AS7263)	Raspberry Pi 4B
SDA	SDA.1 (Pin3)
SCL	SCL.1(Pin5)
VIN	5V (Pin4)
GND	GND (Pin6)
LED control (J2)	
1	GPIO.29 (Pin40)
2	GND (Pin39)
UPS(Lithium battery)	
5V	5V
SDA	SDA.1 (Pin3)
SCL	SCL.1(Pin5)

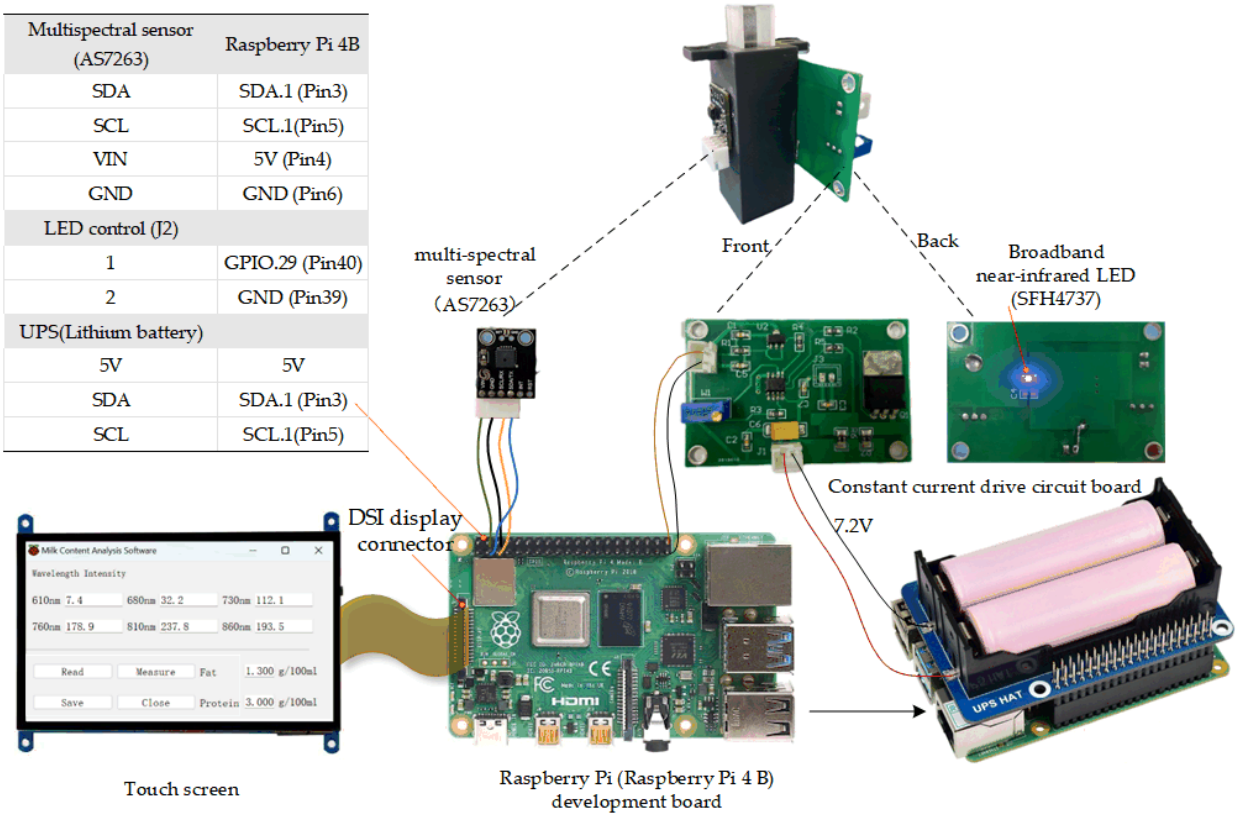

Figure 1. The overall architecture of the portal spectral instrument.

2.2.1. Raspberry Pi

The Raspberry Pi 4 B is a small but powerful minicomputer with a rich interface and flexible operating system serving as the hardware for developing portable instruments. It has support for Wi-Fi, Python, C, and many other programming languages. The size of the module is only 65 mm × 30 mm; it weighs approximately 10 g and has very low power consumption. There is also a DSI interface for connection to the Raspberry Pi's dedicated touchscreen.

2.2.2. Multi-Spectral Sensor

Figure 2 shows the operating principle of the AS7263 circuit. U1 is the AS7263 multi-spectral sensor. U2 is a flash memory chip that stores the firmware of multi-spectral sensors. U3 is a low-dropout voltage regulator chip that converts the 5 V voltage from the Raspberry Pi module to the 3.3 V required for the stable operation of the sensor. AS7263 is a 6-channel spectral sensor that is used for spectral assessment employing NIR light. It consists of six independent optical filters with spectral responses that are defined in the NIR wavelength from approximately 600 nm to 870 nm, with a full width at half maximum of 20 nm. It is highly accurate, stable, and independent of the time of use and temperature.

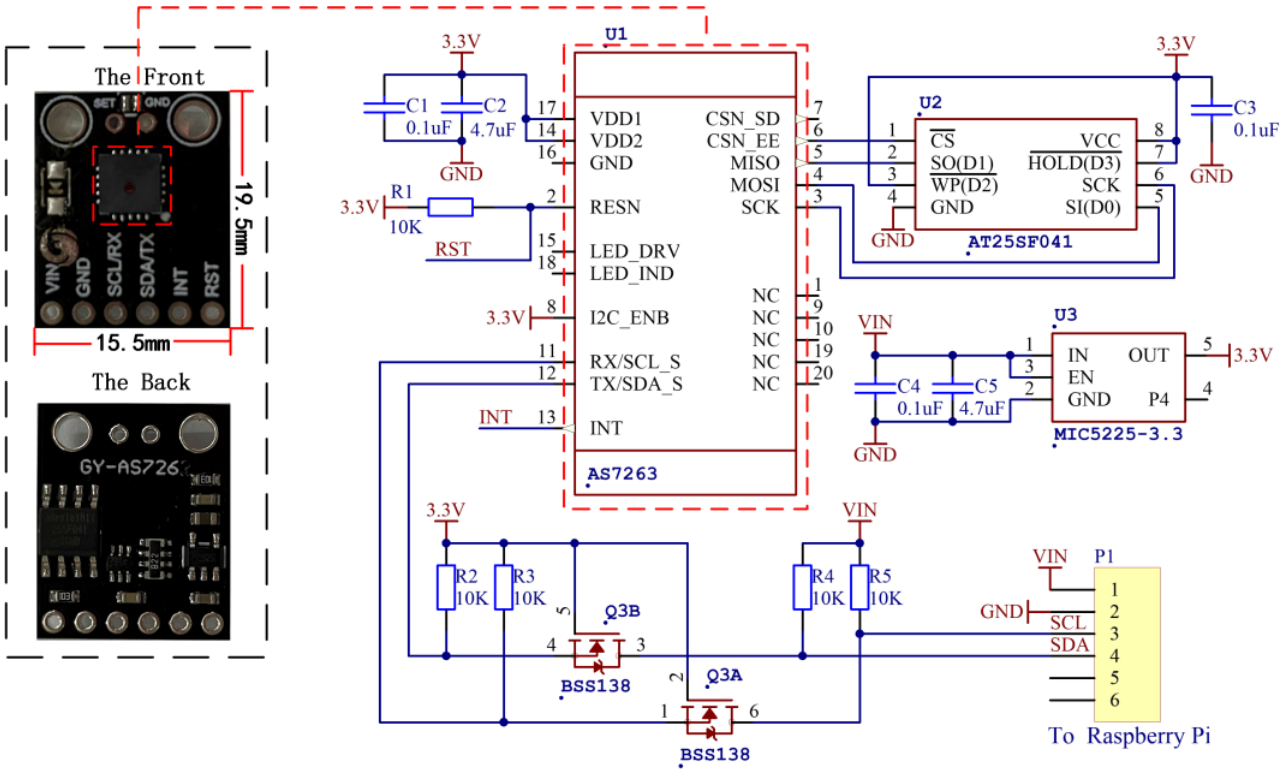

Figure 2. Multi-spectral sensor circuit diagram. The dark blue solid wire is the direct connecting wire of the component; The dark red solid line frame represents the chip; The red dashed box represents the multispectral sensor and its corresponding physical object; The black dashed box represents the multispectral sensor module.

2.2.3. Light Source and Constant-Current Driving Circuit

The constant-current driver circuit of the light source is shown in Figure 3. J1 is used to connect lithium batteries and directly provide a working voltage of 7.2 V for the constant-current driver circuit, and J2 achieves the LED control. D1 is a broadband NIR LED. Potentiometer W1 is used to adjust the current level of D1. The Raspberry Pi's GPIO pin 29 is set to a high value for each measurement. Using J2, the voltage regulator TL431 outputs a 2.5 V reference voltage. After 5 s of D1 conduction, the GPIO pin outputs a low-

level voltage, and D1 turns off. For each conduction time of 5 s, the LED heat generation is small, which improves the service life of the LED and reduces power consumption by using only the copper foil on the back of the PCB to dissipate heat.

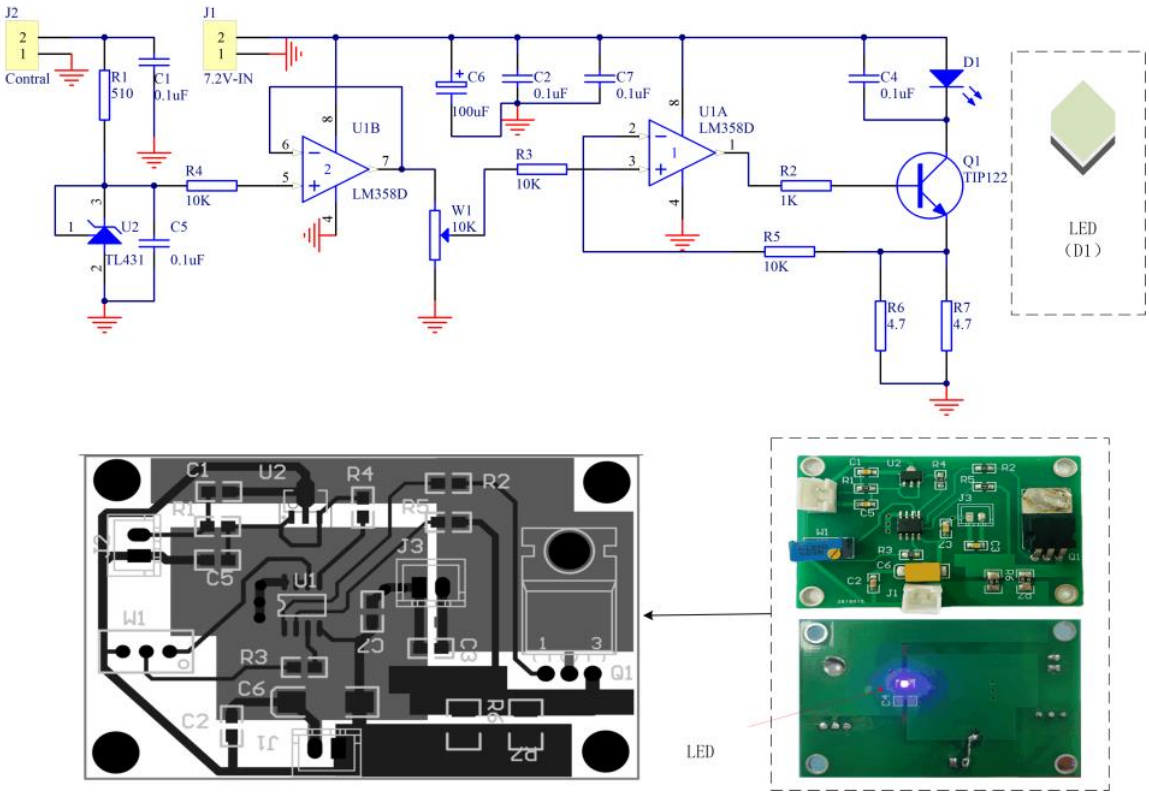

Figure 3. Light source and constant-current driving circuit.

The model of the broadband NIR LED is SFH4737, which is an infrared emitting source designed by Osram Opto Semiconductors (Regensburg, Germany) to support spectral emission in the range of 650–1050 nm, with a rated current of approximately 350 mA.

2.2.4. Power Supply Module

The Li-ion battery module was purchased from Microsnow Electronics (https://www.waveshare.net/wiki/UPS_HAT) (accessed on 1 September 2023) model UPS HAT, Shenzhen City, Guangdong Province. It is a power module designed specifically for the Raspberry Pi 4 B. The 5 V output from the Li-ion battery module is used to power the Raspberry Pi module, whereas the light source and constant-current circuit are directly powered by the 7.2 V Li-ion battery.

2.2.5. Software Setup

Figure 4 shows the milk analysis software (Version 1.2) used in this study. The software was written using Python and pyqt5 code. The Read button reads the wavelength of the milk sample, the Measure button predicts the fat and protein content, the Save button saves this measurement and the predicted information, and the Close button exits the milk composition analysis software.

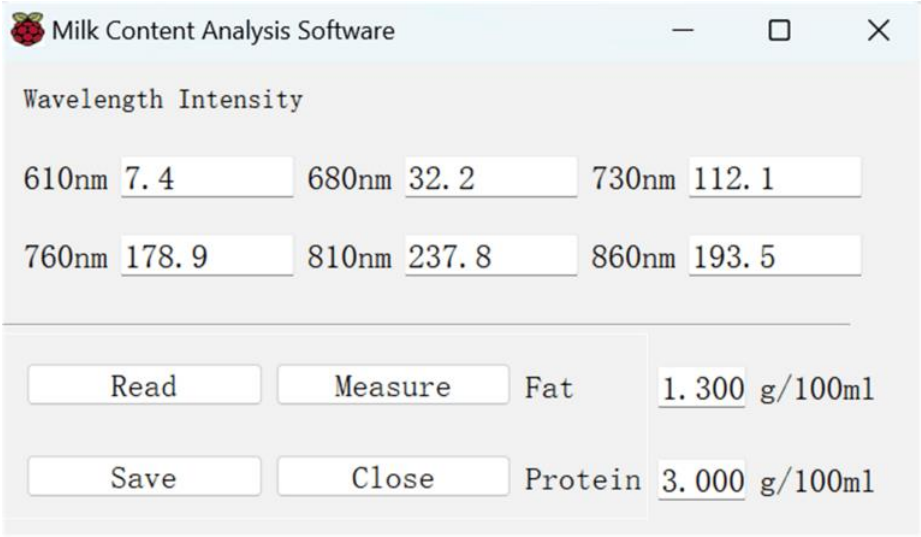

Figure 4. User interface for milk composition measurements.

2.2.6. Operation Instructions

Figure 5 shows the flowchart of the portable instrument used to detect the main milk components. First, 5 mL of a milk sample is measured and placed in a cuvette. The cuvette is then placed in the measuring device. After the instrument is powered on and the Read button is clicked, the LED source is turned on continuously for five seconds. Infrared light is then transmitted through the milk sample, and the multi-spectral sensor acquires six channels of wavelength data as the input features of the model. Finally, the Measure button is clicked, and the model outputs the protein and fat content data.

Figure 5. Detection process.

2.3. Measurement Methods

The Beer-Lambert law [27] is a basic law of spectrophotometry and is used to describe the relationship between the absorbance of a species at a given wavelength and the concentration and thickness of the absorbing species. The use of NIR light for the determination of milk composition is consistent with the Beer-Lambert law:

$$A = -lgT = \lg\left(\frac{I_0}{I}\right) = \sum_{i}^{n} d\varepsilon c, \tag{1}$$

where A, T, I_0, and I denote the absorbance, transmittance, incident light intensity, and transmitted light intensity, respectively; d, ε, and c denote the absorbing layer thickness, molar absorbance coefficient, and concentration of the solution, respectively.

A milk solution is a scatterer, and the extraction of spectral information is complicated. According to the Beer-Lambert law, the spectral characteristics of milk change with its content. Figure 6 shows a schematic of the multi-wavelength data acquisition process.

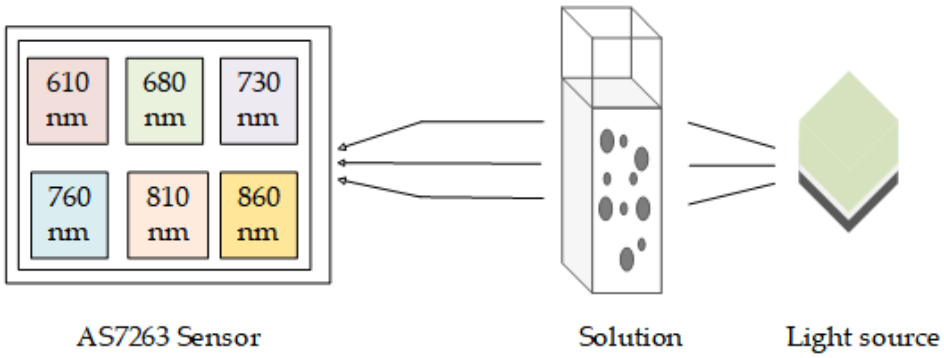

Figure 6. Schematic of the multi-wavelength data acquisition process.

2.4. XGBoost Algorithms

A complex nonlinear relationship exists between milk composition and the corresponding spectral data, and the model trained using the XGBoost machine learning algorithm can fit this nonlinear relationship flexibly and facilitate accurate predictions. The XGBoost algorithm [28] is a gradient-boosting-based machine learning algorithm that builds a stronger model by integrating several weak models to solve predictive regression problems. It has several advantages compared to other machine learning models, including high prediction accuracy, high speed, and good generalization ability.

The model is continuously iterated, and each iteration generates a weak learner. The model is trained based on the residuals of the previous round, which facilitates the continuous improvement of prediction accuracy. Finally, all weak learner outputs are accumulated to build a more accurate model.

The objective function of the XGBoost algorithms is given as follows:

$$Obj = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{t}^{T} \Omega(f_t), \tag{2}$$

where $l(y_i,\hat{y}_i)$ is the loss function, y_i is the true value, $\hat{y}_i$ is the predicted value, $\Omega(f_t)$ is the regularization term, T is the number of prediction trees, n is the number of samples, and Obj is the objective function.

Training error:

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \tag{3}$$

Regularization term:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{t=1}^{T} {\omega_j}^2, \tag{4}$$

where γ is the model's complexity variable, λ is the regular term parameter, and ω_j is the weight of leaf node j. Expanding the objective function using Taylor's formula and setting the derivation as equal to 0, the optimal weight is obtained as follows:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}, G_j = \sum g_i, H_j = \sum h_i, \tag{5}$$

where g_i and h_i are the first- and second-order derivatives of the objective function, respectively. The objective function can be written as:

$$Obj = \frac{1}{2}\sum_{j=1}^{T} \frac{G_j}{H_j + \lambda} + \gamma T \tag{6}$$

2.5. Model Training

This section outlines the training of the milk composition prediction model, as shown in Figure 7. Based on the following experiments conducted on a personal computer, a predictive model for protein and fat in milk was successfully built using machine learning algorithms. The accuracy of the model was verified after parameter tuning and performance tests. It was successfully ported to a Raspberry Pi module, enabling real-time compositional analysis of the embedded system.

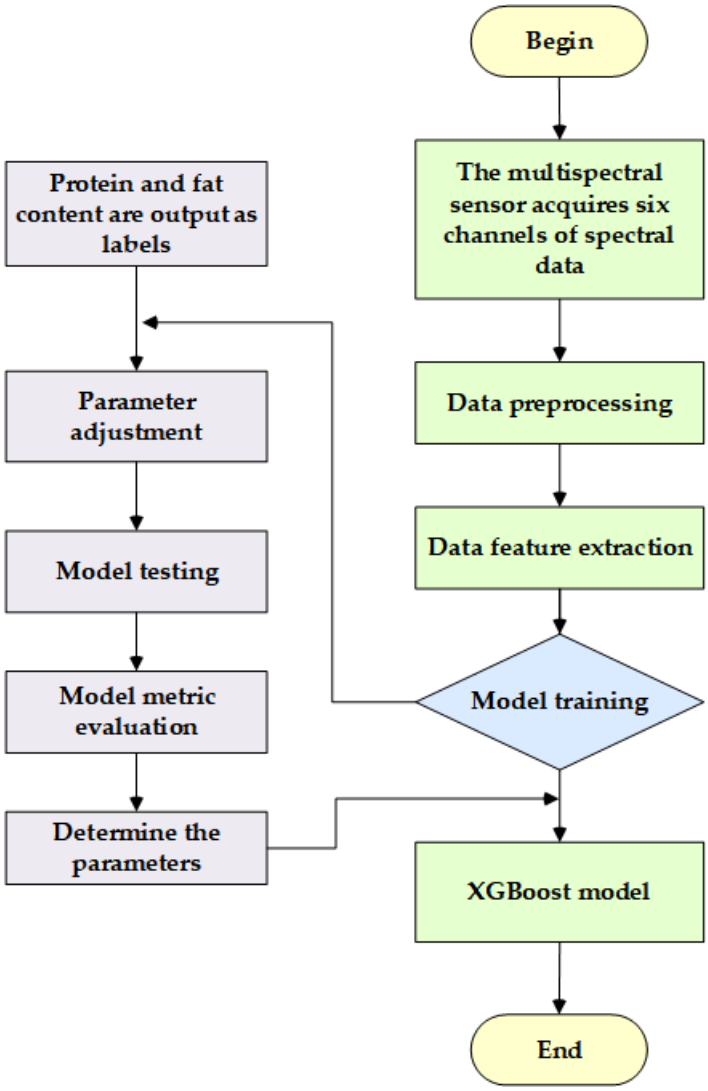

Figure 7. XGBoost model training framework.

2.5.1. Correlation Analysis of Model Characteristics

Based on the data in training set 1, correlation analysis was performed using the six input and output features of the trained XGBoost model. First, the input features and target variables were extracted from the dataset, and the importance of the model input variables was determined and ranked. In the protein prediction model, the 860 nm channel had the highest relative importance of 0.32249, compared to the highest correlation between the 680 nm channel and fat content of 0.95566 in the fat prediction model.

2.5.2. Model Parameter Setting

Before training the XGBoost model, the experimental data were preprocessed to remove redundant data. The parameters were continuously adjusted during the model training process to optimize the evaluation metrics. The parameters that affected the performance of the model were max_depth, n_estimators, learning_rate, subsamples, reg_alpha, and reg_lambda. The values obtained after tuning the reference are listed in Table 1.

Table 1. Parameter values of XGBoost model.

	Parameters	Protein Model	Fat Model
XGBoost	max_depth	6	8
	n_estimators	235	220
	learning_rate	0.14	0.20
	subsample	0.6	0.8
	reg_alpha	0	0
	reg_lambda	0	1

The XGBoost model was mainly evaluated in terms of three evaluation indicators: coefficient of determination (R^2), mean absolute error (*MAE*), and mean square error (*MSE*). They are calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y})^2} \tag{7}$$

$$MAE = \frac{1}{m}\sum_{i}^{m}|y_i - \overline{y}_i| \tag{8}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{9}$$

For the aforementioned indices, R^2 is used to evaluate the degree of model fit. The closer the value of R^2 is to one, the better the model fits. The smaller the *MAE* value, the smaller the prediction error. The smaller the *MSE* value, the higher the accuracy of the model for analyzing experimental data.

As shown in Table 2, the evaluation metrics of the XGBoost model improved after parameter optimization. In the protein model, R^2 increased by 6.11%, whereas *MAE* and *MSE* decreased by 81.26% and 75.51%, respectively. R^2 also increased in the fat model, but the change was relatively small compared to the protein model at 0.80%, corresponding to decreases in mean absolute error (*MAE*) and mean square error (*MSE*) of 84.84% and 78.57%, respectively. The changes in the evaluation indices before and after tuning of the XGBoost model indicated that the tuned XGBoost model performed well.

Table 2. Comparison of XGBoost model before and after tuning.

		R^2	*MAE*	*MSE*
Original model	Protein	0.9251	0.0459	0.0049
	Fat	0.9899	0.0521	0.0070
Optimized model	Protein	0.9816	0.0086	0.0012
	Fat	0.9978	0.0079	0.0015
Improvement	Protein	6.11%	81.26%	75.51%
	Fat	0.80%	84.84%	78.57%

2.5.3. Repeat Verification

The 100 iterations of validation yielded high reliability. In the process, different datasets were randomly used for training, and an average value was used to evaluate the performance of the model. The mean values of R^2, *MAE*, and *MSE* for the protein model were 0.9825, 0.0084, and 0.0012, respectively. The variations in these evaluation indices indicated relatively accurate predictions. The mean R^2, *MAE*, and *MSE* values for the fat model were 0.9981, 0.0070, and 0.0013, respectively. Compared to the protein model, there were individual indicators with relatively large values but within the margin of error. Thus, the prediction model was repeatedly validated to demonstrate that the XGBoost model has good learning performance and can be used for the measurement of protein and fat content in milk.

2.5.4. Five-Fold Cross-Validation

Five-fold cross-validation is a common evaluation method for machine learning models that can reduce the bias of evaluation results owing to the different methods of dividing datasets to yield more reliable results. Table 3 shows that in the five-fold cross-validation test, the mean R^2, *MAE*, and *MSE* values for the protein model were 0.8677, 0.0301, and 0.0088, respectively. The mean determination (R^2), mean absolute error (*MAE*), and mean square error (*MSE*) values for the fat model were 0.9713, 0.0357, and 0.0158, respectively. From the evaluation indices, the XGBoost milk component prediction model was found to be more reliable, with better fitting and higher measurement accuracy.

Table 3. Five-fold cross-validation.

Data	Protein			Fat		
	R^2	*MAE*	*MSE*	R^2	*MAE*	*MSE*
Fold-1	0.8641	0.0325	0.0084	0.9831	0.0354	0.0127
Fold-2	0.9021	0.0238	0.0052	0.9209	0.0413	0.0319
Fold-3	0.7832	0.0395	0.0163	0.9787	0.0343	0.0135
Fold-4	0.8491	0.0307	0.0102	0.9903	0.0312	0.0078
Fold-5	0.9401	0.0244	0.0042	0.9837	0.0363	0.0132
Average	0.8677	0.0301	0.0088	0.9713	0.0357	0.0158

3. Results

3.1. Model Comparison

To evaluate the detection performance of the XGBoost model, the following algorithms were selected for comparison: (1) Linear Regression [29] (LR), a classical machine learning algorithm for solving regression problems; (2) Stochastic Gradient Descent [30] (SGD), an optimization algorithm for minimizing the loss function (it is a variant of gradient descent algorithms, which gradually reduce the value of the loss function by iteratively updating the model parameters); (3) Multilayer Perceptron [31] (MLP), a neural network-based machine learning model that is often used to solve classification and regression problems; (4) Gradient Boosted Regression Tree [32] (GBRT), an integrated tree-based learning algorithm that accumulates the results of all regression tree outputs; (5) Random Forest [33] (RF), an integrated learning algorithm for solving classification and regression problems based on an integrated approach of decision trees, which constructs multiple decision trees and combines their predictions to perform classification or regression.

Table 4 presents the results of the performance analysis of the same dataset using the stated models. The XGBoost model achieved a better learning rate and lower error than the other machine learning algorithms, as well as accurate estimates of the protein and fat in the milk samples.

Table 4. Comparison of the performance of different models.

Method		R^2	*MAE*	*MSE*
LR	Protein	0.1809	0.1806	0.0541
	Fat	0.8624	0.2290	0.0946
SGD	Protein	0.1165	0.1866	0.0584
	Fat	0.8112	0.2739	0.1300
MLP	Protein	0.6405	0.1155	0.0237
	Fat	0.5792	0.1270	0.0278
GBRT	Protein	0.7905	0.0847	0.0134
	Fat	0.9919	0.0959	0.0160
RF	Protein	0.9703	0.0136	0.0020
	Fat	0.9951	0.0172	0.0035
XGBOOST	Protein	0.9864	0.0048	0.0009
	Fat	0.9994	0.0079	0.0013

3.2. Instrument Validation Results

Figure 8 shows the results for Test Set 1. The MAE and MSE of the protein model illustrated in Figure 8a were 0.0086 and 0.0012, respectively. The *MAE* and *MSE* of the fat model illustrated in Figure 8b were 0.0077 and 0.0015, respectively. There were still individual samples with large differences between the true and predicted values, which may have been induced by the measurement process. The overall forecast had high accuracy and a relatively small error. The XGBoost model showed better generalization in the case of the unknown milk sample data used in the training process compared to the other models.

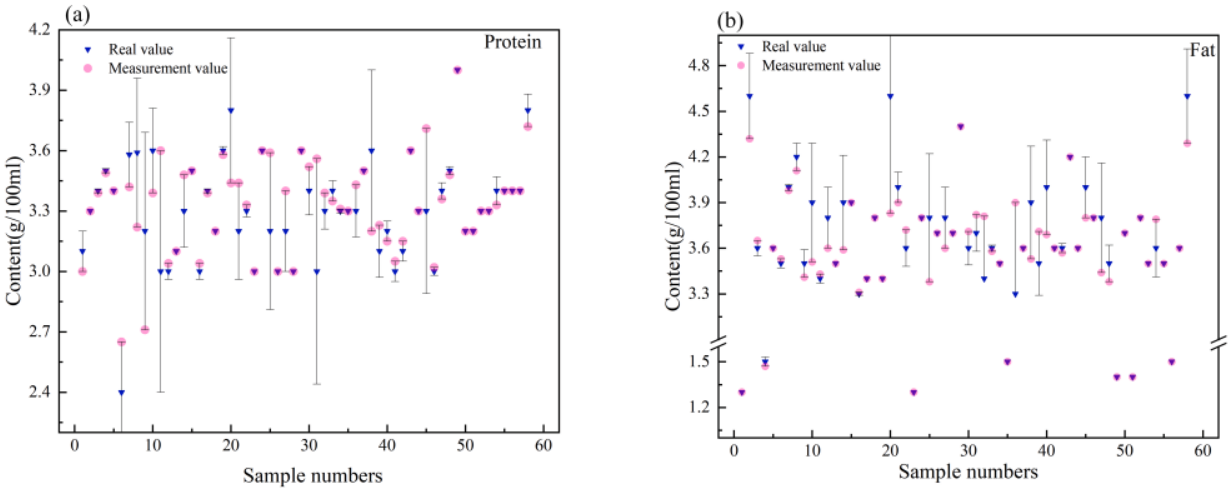

Figure 8. Protein model prediction results for Test Set 1 (**a**) and fat model prediction results for Test Set 1 (**b**).

However, the model did not learn all the milk samples. To further test the predictive ability of the model for new samples, two milk samples, Milk-1 and Milk-2, from Test Set 2 were used. The results are shown in Figure 9. The dataset contains 40 unlearned milk samples with an actual protein content of 3.3 and 3.5 g/100 mL and an actual fat content of 4.6 and 1.9 g/100 mL for the two milk samples.

The protein model illustrated in Figure 9a predicted more accurate results for Milk-1 and Milk-2. The average predicted values were very close to the mean of the actual values of 3.3004 and 3.4976 g/100 mL, respectively. The *MAE* and *MSE* of the model were 0.0660 and 0.0109, respectively. This indicates that the average prediction error of the model was relatively small, and the difference between this value and the true value was small.

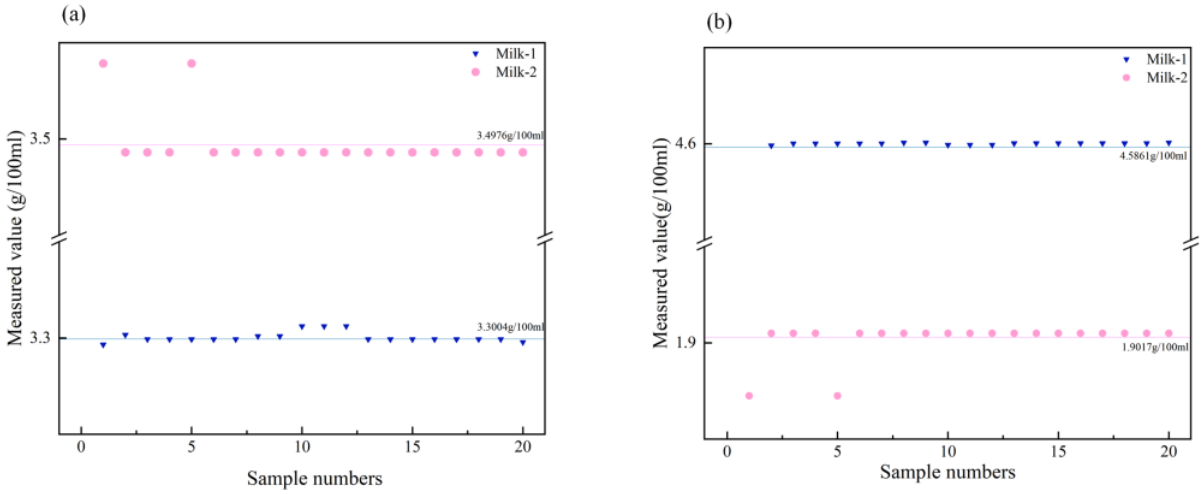

Figure 9. Protein model prediction results for Test Set 2 (**a**) and fat model prediction results for Test Set 2 (**b**).

The prediction results of the fat model illustrated in Figure 9b were more accurate for Milk-1 and Milk-2, with mean prediction values of 4.5861 and 1.9017 g/100 mL, respectively. The *MAE* and *MSE* of the model were 0.0580 and 0.0072, respectively. This indicates that the average prediction error of the model and the difference between this value and the true value were small.

The mean error of the two milk samples was relatively low, indicating good generalization to the new samples and good accuracy of the XGBoost measurement model. By testing the samples and comparing the data, our designed detection instrument can achieve the expected accuracy.

4. Discussion

4.1. Near-Infrared Spectroscopy Analysis Using Machine Learning

The objective of this study was to improve the performance of traditional NIR milk composition detection methods through machine learning algorithms and sensor technology. Infrared spectroscopy, which is the method for detecting milk components, is fast and capable of multi-component analysis and non-destructive testing. The NIR wavelength is generally selected in the range of 400–1000 nm [34–36], and wavelength data are generally acquired through a spectrometer to obtain spectral data over the entire wavelength range. However, the instrument structure using spectrometer technology is complex and expensive. The portable milk composition detection instrument uses several typical NIR wavelengths for milk detection, requiring multiple infrared receiving units combined with a single wavelength filter to achieve multi-channel infrared measurement. The new type of multi-wavelength sensor used in this article simultaneously obtains wavelength data from six channels corresponding to representative wavelengths (610, 680, 730, 760, 810, 860 nm) in the range of 400–1000 nm, which cover the infrared characteristics of protein and fat in milk. Thus, an NIR detection method for milk is proposed.

In previous studies, the partial least-squares method was used for regression of spectral data to measure the fat and protein contents in milk [36–38]. The coefficients of determination (R^2) of better protein and fat prediction models were 0.974 and 0.973, respectively [36]. We use the machine learning algorithm XGBoost [28], which is a powerful ensemble learning algorithm that can better fit the nonlinear relationship between wavelengths and components, significantly reduce the model bias, and increase the model accuracy. With sufficient training with a large quantity of sample data and adjustment of model parameters, the prediction model exhibited a good generalization ability and no overfitting [39], indicating that the measurement results were accurate.

4.2. Advantages of Proposed Method

The use of multi-wavelength sensors simplifies optical system design, multi-channel sensor design, and optical signal acquisition and processing. The new sensor technology is not affected by usage time and temperature and has extremely high accuracy and stability. The collection of multiple wavelengths enables a milk sample to correspond to multiple wavelength data, enabling the application of machine learning algorithms for sample composition measurement. Machine learning algorithms can flexibly fit the nonlinear relationship between milk wavelength and component content, with high measurement accuracy, fast running speed, and good generalization ability. The application of broadband NIR LED (SFH4737) can support the emission spectrum range of NIR) from 650 nm to 1050 nm, which can replace conventional NIR light sources such as incandescent lamps, miniature tungsten halide lamps, and tritium lamps. The light source has good stability, long lifespan, and high energy, and the small size design also increases the portability of the detection instrument. The adoption of machine learning algorithms by embedded systems has increased the intelligence of devices without the need for powerful computing devices, reduced the demand for cloud computing, and reduced the cost of portable instruments.

4.3. Potential Interference

In the process of measuring milk composition, there are various potential interference factors that may affect the accuracy, such as light scattering [40], environmental temperature changes [41], sample bacterial infections [42], environmental light, and the material and fixation of colorimetric plates.

Scattering interference is a common problem in the measurement of milk composition, and the propagation of light in milk is very complex. When light shines on milk, scattering occurs on the surface of protein and fat particles. To reduce the impact of the aforementioned interference factors, a new multi-wavelength sensor with temperature compensation and strong penetration was adopted. It provides six channels of wavelength data for milk samples. Calibration models were established between wavelength and protein fat through machine learning algorithms, effectively eliminating the influence of scattering and improving model prediction accuracy.

Temperature changes in the environment can cause instability of the light source, and the scattering coefficient decreases with increasing temperature. The experiments in this study were conducted at room temperature (25 °C), which had little impact on the experimental results. In the next step of research, we will consider adding different temperatures as new features to machine model training to resist the influence of temperature on the experiment.

When excessive external bacteria are mixed with milk samples [43], they decompose the protein and fat in the milk, causing measurement errors, which cannot be completely avoided. The material of the cuvette can affect the refraction of light. In this study, a glass cuvette was used, and the measurement error in the visible light region can be ignored. The colorimetric dish was fixed vertically through a spring lock to ensure that light shined vertically onto the transparent surface.

5. Conclusions

This article applies embedded technology, multi-wavelength sensors, and machine learning algorithms to develop a relatively portable instrument for detecting the main components of milk. We designed a constant-current driving circuit and wavelength acquisition module for broadband NIR LED light sources, established an optimal prediction model, and transplanted the model into an embedded system. The model results indicate that the MSE of the protein and fat models is 0.12% and 0.15%, respectively. It has good measurement performance for protein and fat content in milk. The milk composition detector is lightweight and low cost, has a small volume, simple operation, fast measurement speed, and stable operation, and it can achieve on-site rapid detection of milk samples.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app132212320/s1, Table S1: milk Sample Data.

Author Contributions: Conceptualization, Y.W. and Q.W.; methodology, Y.W.; software, K.Z.; validation, Y.W., S.S. and K.Z.; data curation, Q.W.; writing—original draft preparation, Y.W.; writing—review and editing, K.Z. and S.S.; visualization, S.S. and Q.W.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the 2022 Graduate Innovation Fund of Huaibei Normal University (No. cx2023044) and the 2022 New Era Education Provincial Quality Engineering Project (Graduate Education) (No. 2022cxcysj115).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/wang3147 (accessed on 10 October 2023).

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

References

1. Hayes, E.; Wallace, D.; O'Donnell, C.; Greene, D.; Hennessy, D.; O'Shea, N.; Tobin, T.J.; Fenelon, M.A. Trend analysis and prediction of seasonal changes in milk composition from a pasture-based dairy research herd. *J. Dairy Sci.* **2023**, *106*, 2326–2337. [CrossRef]
2. Shim, H.W.; Lee, W.Y.; Kim, H.W.; Park, J.K.; Cho, K.; Yeo, J.M.; Park, H.J. Physiological Effects of Hydrolyzed Skim Milk and Probiotics on Osteoporosis Models. *Appl. Sci.* **2023**, *13*, 10424. [CrossRef]
3. Willett, W.C.; Ludwig, D.S. Milk and health. *N. Engl. J. Med.* **2020**, *382*, 644–654. [CrossRef] [PubMed]
4. Kumar, D.N.; Pinker, N.; Shtenberg, G. Porous silicon Fabry–Pérot interferometer for N-acetyl-β-d-glucosaminidase biomarker monitoring. *ACS Sens.* **2020**, *5*, 1969–1976. [CrossRef] [PubMed]
5. Di Marzo, L.; Pranata, J.; Barbano, D.M. Measurement of casein in milk by Kjeldahl and sodium dodecyl sulfate–polyacrylamide gel electrophoresis. *J. Dairy Sci.* **2021**, *104*, 7448–7456. [CrossRef]
6. Gurd, C.; Jefferson, B.; Villa, R.; De Castro Rodriguez, C. Determination of fats, oils and greases in food service establishment wastewater using a modification of the Gerber method. *Water Environ. J.* **2020**, *34*, 5–13. [CrossRef]
7. Di Stefano, V.; Avellone, G.; Bongiorno, D.; Cunsolo, V.; Muccilli, V.; Sforza, S.; Dossena, A.; Drahos, L.; Vékey, K. Applications of liquid chromatography-mass spectrometry for food analysis. *J. Chromatogr. A* **2012**, *1259*, 74–85. [CrossRef] [PubMed]
8. Eskildsen, C.E.; Skov, T.; Hansen, M.S.; Larsen, L.B.; Poulsen, N.A. Quantification of bovine milk protein composition and coagulation properties using infrared spectroscopy and chemometrics: A result of collinearity among reference variables. *J. Dairy Sci.* **2016**, *99*, 8178–8186. [CrossRef]
9. 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]
10. Soyeurt, H.; Dehareng, F.; Gengler, N.; McParland, S.; Wall, E.P.B.D.; Berry, D.P.; Coffey, M.; Dardenne, P. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *J. Dairy Sci.* **2011**, *94*, 1657–1667. [CrossRef]
11. Bonfatti, V.; Vicario, D.; Lugo, A.; Carnier, P. Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle. *J. Dairy Sci.* **2017**, *100*, 5526–5540. [CrossRef]
12. Dabrowska, A.; David, M.; Freitag, S.; Andrews, A.M.; Strasser, G.; Hinkov, B.; Schwaighofer, A.; Lendl, B. Broadband laser-based mid-infrared spectroscopy employing a quantum cascade detector for milk protein analysis. *Sens. Actuators B Chem.* **2022**, *350*, 130873. [CrossRef]
13. Zhao, X.; Song, Y.; Zhang, Y.; Cai, G.; Xue, G.; Liu, Y.; Chen, K.; Zhang, F.; Wang, K.; Zhang, M.; et al. Predictions of milk fatty acid contents by mid-infrared spectroscopy in Chinese Holstein cows. *Molecules* **2023**, *28*, 666. [CrossRef]
14. 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]
15. De Marchi, M.; Penasa, M.; Zidi, A.; Manuelian, C.L. Invited review: Use of infrared technologies for the assessment of dairy products—Applications and perspectives. *J. Dairy Sci.* **2018**, *101*, 10589–10604. [CrossRef]
16. Yakubu, H.G.; Kovacs, Z.; Toth, T.; Bazar, G. The recent advances of near-infrared spectroscopy in dairy production—A review. *Crit. Rev. Food Sci. Nutr.* **2022**, *62*, 810–831. [CrossRef] [PubMed]
17. Mancini, M.; Mazzoni, L.; Gagliardi, F.; Balducci, F.; Duca, D.; Toscano, G.; Mezzetti, B.; Capocasa, F. Application of the non-destructive NIR technique for the evaluation of strawberry fruits quality parameters. *Foods* **2020**, *9*, 441. [CrossRef] [PubMed]
18. Kapse, S.; Kedia, P.; Kausley, S.; Rai, B. Nondestructive Evaluation of Banana Maturity Using NIR AS7263 Sensor. *J. Nondestruct. Eval.* **2023**, *42*, 30. [CrossRef]

19. Li, J.; Wang, Q.; Xu, L.; Tian, X.; Xia, Y.; Fan, S. Comparison and optimization of models for determination of sugar content in pear by portable Vis-NIR spectroscopy coupled with wavelength selection algorithm. *Food Anal. Methods* **2019**, *12*, 12–22. [CrossRef]
20. Coppa, M.; Ferlay, A.; Leroux, C.; Jestin, M.; Chilliard, Y.; Martin, B.; Andueza, D. Prediction of milk fatty acid composition by near infrared reflectance spectroscopy. *Int. Dairy J.* **2010**, *20*, 182–189. [CrossRef]
21. Uusitalo, S.; Diaz-Olivares, J.; Sumen, J.; Hietala, E.; Adriaens, I.; Saeys, W.; Utriainen, M.; Frondelius, L.; Pastell, M.; Aernouts, B. Evaluation of MEMS NIR Spectrometers for On-Farm Analysis of Raw Milk Composition. *Foods* **2021**, *10*, 2686. [CrossRef]
22. Saranwong, S.; Kawano, S. System design for non-destructive near infrared analyses of chemical components and total aerobic bacteria count of raw milk. *J. Near Infrared Spectrosc.* **2007**, *16*, 389–398. [CrossRef]
23. Muñoz-Salinas, F.; Andrade-Montemayor, H.M.; De la Torre-Carbot, K.; Duarte-Vázquez, M.Á.; Silva-Jarquin, J.C. Comparative analysis of the protein composition of goat milk from French Alpine, Nubian, and Creole breeds and Holstein Friesian cow milk: Implications for early infant nutrition. *Animals* **2022**, *12*, 2236. [CrossRef] [PubMed]
24. Zaky, Z.A.; Sharma, A.; Alamri, S.; Saleh, N.; Aly, A.H. Detection of fat concentration in milk using ternary photonic crystal. *Silicon* **2021**, *14*, 6063–6073. [CrossRef]
25. Almawgani, A.H.; Daher, M.G.; Taya, S.A.; Mashagbeh, M.; Colak, I. Optical detection of fat concentration in milk using MXene-based surface plasmon resonance structure. *Biosensors* **2022**, *12*, 535. [CrossRef]
26. Yang, B.; Zhu, Z.; Gao, M.; Yan, X.; Zhu, X.; Guo, W. A portable detector on main compositions of raw and homogenized milk. *Comput. Electron. Agric.* **2020**, *177*, 105668. [CrossRef]
27. Xingcai, L.; Kun, N. Effectively predict the solar radiation transmittance of dusty photovoltaic panels through Lambert-Beer law. *Renew. Energy* **2018**, *123*, 634–638. [CrossRef]
28. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
29. Maulud, D.; Abdulazeez, A.M. A review on linear regression comprehensive in machine learning. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 140–147. [CrossRef]
30. Ketkar, N.; Ketkar, N. Stochastic gradient descent. In *Deep Learning with Python: A Hands-on Introduction*; Apress: Berkeley, CA, USA, 2017; pp. 113–132.
31. Rozos, E.; Dimitriadis, P.; Mazi, K.; Koussis, A.D. A multilayer perceptron model for stochastic synthesis. *Hydrology* **2021**, *8*, 67. [CrossRef]
32. Wang, L.; Zhang, Y.; Yao, Y.; Xiao, Z.; Shang, K.; Guo, X.; Yang, J.; Xue, S.; Wang, J. Gbrt-based estimation of terrestrial latent heat flux in the haihe river basin from satellite and reanalysis datasets. *Remote Sens.* **2021**, *13*, 1054. [CrossRef]
33. Rigatti, S.J. Random forest. *J. Insur. Med.* **2017**, *47*, 31–39. [CrossRef] [PubMed]
34. Niero, G.; Penasa, M.; Gottardo, P.; Cassandro, M.; De Marchi, M. Selecting the most informative mid-infrared spectra wavenumbers to improve the accuracy of prediction models for detailed milk protein content. *J. Dairy Sci.* **2016**, *99*, 1853–1858. [CrossRef] [PubMed]
35. Kawamura, S.; Kawasaki, M.; Nakatsuji, H.; Natsuga, M. Near-infrared spectroscopic sensing system for online monitoring of milk quality during milking. *Sens. Instrum. Food Qual. Saf.* **2007**, *1*, 37–43. [CrossRef]
36. Kucheryavskiy, S.; Melenteva, A.; Bogomolov, A. Determination of fat and total protein content in milk using conventional digital imaging. *Talanta* **2014**, *121*, 144–152. [CrossRef]
37. Gastélum-Barrios, A.; Soto-Zarazúa, G.M.; Escamilla-García, A.; Toledano-Ayala, M.; Macías-Bobadilla, G.; Jauregui-Vazquez, D. Optical Methods Based on Ultraviolet, Visible, and Near-Infrared Spectra to Estimate Fat and Protein in Raw Milk: A Review. *Sensors* **2020**, *20*, 3356. [CrossRef]
38. Dos Santos Pereira, E.V.; de Sousa Fernandes, D.D.; 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]
39. Jabbar, H.; Khan, R.Z. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Comput. Sci. Commun. Instrum. Devices* **2015**, *70*, 978–981.
40. Aernouts, B.; Van Beers, R.; Watté, R.; Huybrechts, T.; Lammertyn, J.; Saeys, W. Visible and near-infrared bulk optical properties of raw milk. *J. Dairy Sci.* **2015**, *98*, 6727–6738. [CrossRef] [PubMed]
41. Zhu, X.; Guo, W.; Jia, Y.; Kang, F. Dielectric properties of raw milk as functions of protein content and temperature. *Food Bioprocess Technol.* **2015**, *8*, 670–680. [CrossRef]
42. O'connell, A.; Ruegg, P.L.; Jordan, K.; O'brien, B.; Gleeson, D. The effect of storage temperature and duration on the microbial quality of bulk tank milk. *J. Dairy Sci.* **2016**, *99*, 3367–3374. [CrossRef] [PubMed]
43. Marchand, S.; De Block, J.; De Jonghe, V.; Coorevits, A.; Heyndrickx, M.; Herman, L. Biofilm formation in milk production and processing environments; influence on milk quality and safety. *Compr. Rev. Food Sci. Food Saf.* **2012**, *11*, 133–147. [CrossRef]