Food Quality and Safety, 2023, **7**, 1–9
https://doi.org/10.1093/fqsafe/fyad014
Advance access publication 7 February 2023
Article

OXFORD

Article

A rapid method of identifying mastitis degrees of bovines based on dielectric spectra of raw milk

Zhuozhuo Zhu (朱卓卓)[1,2,†], Biying Lin (林碧莹)[1,†], Xinhua Zhu (朱新华)[1,3,*], and Wenchuan Guo (郭文川)[1,4,*]

[1]College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, China
[2]School of Physics and Electronic Information, Yan'an University, Yan'an, China
[3]Shaanxi Research Center of Agricultural Equipment Engineering Technology, Yangling, China
[4]Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, China.

[†]These two authors contributed equally to this work.
**Correspondence to*: Xinhua Zhu, College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling 712100, China. E-mail: zxh920@nwsuaf.edu.cn; Wenchuan Guo, College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling 712100, China. E-mail: wencguo@nwafu.edu.cn, guowenchuan69@126.com

Abstract

Bovine mastitis is the most complex and costly disease in the dairy industry worldwide. Somatic cell count (SCC) is accepted as an international standard for diagnosing mastitis in cows, but most instruments used to detect SCC are expensive, or the detection speed is very low. To develop a rapid method for identifying mastitis degree, the dielectric spectra of 301 raw milk samples at three mastitis grades, i.e., negative, weakly positive, and positive grades based on SCC, were obtained in the frequency range of 20–4500 MHz using coaxial probe technology. Variable importance in the projection method was used to select characteristic variables, and principal component analysis (PCA) and partial least squares (PLS) were used to reduce data dimension. Linear discriminant analysis, support vector classification (SVC), and feed-forward neural network models were established to predict the mastitis degrees of cows based on 22 principal components and 24 latent variables obtained by PCA and PLS, respectively. The results showed that the SVC model with PCA had the best classification performance with an accuracy rate of 95.8% for the prediction set. The research indicates that dielectric spectroscopy technology has great potential in developing a rapid detector to diagnose mastitis in cows *in situ* or online.

Keywords: Mastitis; somatic cell count; dielectric spectra; qualitative analysis.

Introduction

Mastitis, or inflammation of the mammary gland, is a complex and costly disease to the dairy industry (Mariani *et al.*, 2022). It results in udder pain and reduces milk quality, yield, and dairy production efficiency. For example, mastitis causes decreases in casein and lactose contents and increases in somatic cell count (SCC), fatty acid, sodium, and chloride values (Ogola *et al.*, 2007; Wickström *et al.*, 2009). A practical solution to control mastitis is treating cows with antibiotics (Jafari *et al.*, 2019). However, antibiotic residual, which can be transferred from milk or dairy products to humans, may be harmful to human health. The sooner mastitis is diagnosed, the less antibiotics need to be used. Therefore, diagnosing mastitis is very important for cows, farms, and consumers.

At present, SCC is accepted as an international standard for diagnosing mastitis in cows (IDF, 2013). In China, a cow is regarded as healthy if the SCC of its produced milk is lower than 500×10^3 cells/mL. Otherwise, the cow is suffering from mastitis. Four grades are used to describe the mastitis degrees of the cow suffering from mastitis, including negative, weakly positive, positive, and strongly positive if the SCC is lower than 500×10^3 cells/mL, between 500×10^3 cells/mL and 1500×10^3 cells/mL, between 1500×10^3 cells/mL and 5000×10^3 cells/mL, and higher than 5000×10^3 cells/mL, respectively (MARA, 2015). The industry-standard methods applied to detect milk SCC and diagnose cows' subclinical mastitis include microscope counting, electronic and particle somatic cell counting, and fluorescent photoelectric somatic cell counting (MARA, 2004). Another rapid detection technique widely used in diagnosing subclinical mastitis for cows is observing formed gels after adding special diagnostic liquids into milk to lyze somatic cells (Schalm and Noorlander, 1957). However, most of these methods have shortcomings of complex operations, long detection time, rough judgment, or are unsuitable for in-line use.

Currently, automated SCC counters are commonly used for rapidly obtaining milk SCC, such as Fossomatic 7 DC (Foss Electric) and Somacount FC (Bentley Instruments Inc., Chaska, MN, USA). High instrument prices and operation costs limit the application range of these SCC counters. Several methods have been studied or developed to quickly detect SCC or mastitis degrees. Bioluminescence

Received 12 November 2022; Revised 23 January 2023; Editorial decision 28 January 2023

(Frundzhyan *et al.*, 2008), viscosity (Atasever *et al.*, 2012), immunofluorescence (Becheva *et al.*, 2018), and colorimetric assay (Thiruvengadam *et al.*, 2020) are very effective has also been studied detection only under severe pretreatment conditions, making them too difficult to use in online detection. Good relationships between SCC and near-infrared spectra of milk have been reported, but the accuracy was influenced by the main constituents of milk (Tsenkova *et al.*, 2001; Melfsen *et al.*, 2012; Iweka *et al.*, 2020). Terahertz spectroscopy has also been studied for SCC detection, and monochromatic terahertz waves have been is found to be highly related to SCC (Naito et al., 2013, 2015). However, low measurement accuracy and high cost of equipment hinder the practical application of terahertz technology.

Electrical conductivity has been used to monitor cow mastitis degrees, but its classification accuracy is low (Norberg *et al.*, 2004; Lien *et al.*, 2016). Grillo *et al.* (2002) found that intact somatic cells in milk influence the dielectric constant of raw cow milk at 300 kHz, but the predicted deviations reached 35% below 22 °C and 20% above 30 °C. Dielectric spectroscopy, a fast, easy technique, potentially capable of in-line detection, has great potential in determining the qualitative characteristics of some agricultural products and foods, such as classifying olive oils during the storage period (Sanaeifar *et al.*, 2018) and predicting total bacteria count of raw goat milk (Zhu *et al.*, 2019). However, no studies have been carried out to qualitatively identify mastitis degrees of cows using dielectric spectroscopy over a wide frequency range. Therefore, the objectives of this study were: (1) to obtain the dielectric spectra of raw cow milk samples with different SCC values over the frequency range of 20–4500 MHz, (2) to build qualitative models for identifying mastitis degrees based on the obtained spectra of milk, and (3) to provide a fast detection method for mastitis degrees of cows.

Materials and methods

Milk samples

When suspected clinical mastitis cases of cows, breed 'Holstein', in three farms located in Yangling, Shannxi, China, were reported by feeders, the milk samples (original mastitis samples) were collected from one or two cows with suspected clinical mastitis at 4 p.m. on the reported days. Meanwhile, nine milk samples (normal samples) were collected from observed healthy cows at the same sampling time. Composite milk sample were collected after removal of the fore-stripping of the first three strips per quarter. The milk samples were transported from the farms to the laboratory in 20 min.

When two mastitis samples were collected in a single sampling, the two original mastitis samples were mixed at ratios of 4:1 and 1:4 to obtain two new mastitis samples. Then each mastitis sample was mixed with a randomly selected milk sample among nine normal samples at ratios of 2:1, 1:2, 1:6, and 1:12. In total, 301 milk samples were obtained in this study, including 35 mastitis samples, 126 normal milk samples, and 140 mixed samples.

Each prepared sample was separated into two aliquots. One aliquot was kept at room temperature (24±1 °C) and used to measure milk dielectric spectra, pH, and electrical conductivity within 6 h after milking. Another aliquot was kept in a refrigerator at 4 °C and applied to measure milk components and SCC within 48 h of milking. To delay the deterioration of milk, 0.03 g potassium dichromate (analytical grade, 99.8%; Tianjin Bodi Chemical Co. Ltd., Tianjin, China) was added to each 50 mL milk sample. To distribute the components evenly, each milk sample was stirred by an electric blender (OA2000, Shanghai Ouhor Equipment Co., Shanghai, China) for approximately 1 min before measurement.

Measurement on milk components and SCC

The components and SCC of all milk samples were determined using a Milk Component Analyzer (Combi 500, Bentley Instruments Inc., Chaska, MN, USA). The analyzer had two probes. One was used to measure milk components, including fat, protein, lactose, and total solids contents, and another was used for SCC measurement based on the flow cytometry technique. Measurements were carried out in triplicate for each sample, and the mean was calculated and used in this study.

Measurement on dielectric spectra

Dielectric constant ε' and dielectric loss factor ε'' are usually used to describe the dielectric properties of food and agricultural products. The former represents energy storage capacity, and the latter reflects energy consumption (Altemimi *et al.*, 2019). These two parameters were obtained by a dielectric properties measurement system that consisted of an Agilent E5071C network analyzer (Agilent Technologies, Penang, Malaysia), an Agilent 85070E open-ended coaxial probe, Agilent N6314A coaxial cable, an Agilent 85070 dielectric probe kit software, a constant temperature water bath (DK-98-1, Tianjin Taisite Instrument Ltd., Tianjin, China), and a computer. Before measurement, the system was preheated for one hour to make the system stable. Then, calibration operations on the network analyzer and coaxial probe were conducted. Detailed information on the measurement system and calibration procedures can be found elsewhere (Guo *et al.*, 2015; Zhang *et al.*, 2016).

Each milk sample was filled in a 50 mL centrifuge tube and placed in a water bath at 25 °C, and a thermometer with 0.1 °C precision was used to check the milk temperature. A laboratory hydraulic elevator was used to elevate the water bath and immerse the probe in the milk. When the milk temperature reached 25 °C, the dielectric properties at 201 discrete frequencies over the frequency range of 20–4500 MHz were obtained. Each sample was measured in triplicate, and the mean was used as the result. Finally, two dielectric spectra, including 201 values of ε' and 201 values of ε'', were obtained for each milk sample.

Each sample's pH and electrical conductivity were measured using a pH meter (PHSJ-3F, Shanghai Precision & Scientific Instrument Co., Ltd., Shanghai, China) and a conductivity meter (DDSJ-308A, Shanghai Precision & Scientific Instrument Co., Ltd., Shanghai, China) at room temperature, respectively. The means of triplicates for pH and electrical conductivity are reported in this study.

Spectra preprocessing and analysis methods

Spectra preprocessing

Appropriate spectra preprocessing is necessary for improving the stability and prediction performance of established models. Savitzky–Golay (SG) smoothing and standard normal variate (SNV) are widely used methods to reduce high-frequency noises and to separate physical and chemical variances in spectral data, respectively (Afseth *et al.*, 2006; Guo *et al.*, 2020), as well as SNV after SG smoothing (SG+SNV), and were applied to preprocess dielectric spectra in this study. Detailed information about SG smoothing and SNV has been previously reported (Savitzky and Golay, 1964; Rinnan *et al.*, 2009).

Outlier detection

Outliers that might dramatically influence the analysis results should be removed before building prediction models (Thennadil *et al.*, 2018). In this study, one-class support vector machine (1SVM) and local outlier factor (LOF) were used as outlier detection methods.

LOF is the first local outlier detection method, and is efficient for high-dimensional datasets (Breunig *et al.*, 2000). The local reachability density of the k-nearest neighbors set of the test sample is compared with that of each number in the k-nearest neighbors set. The sample with a small local reachability density relative to its nearest neighbors could be an outlier.

1SVM, proposed by Schölkopf *et al.* (2001), is also a popularly used outlier detection method. In this method, the image vectors in the feature space are separated from the origin by a hyperplane with the largest possible margin. The vectors in the half-space containing the origin are marked as outliers.

Sample division

A proper sample division method could choose sufficient representative samples for training and verifying models and promote the efficiency and accuracy of modeling. The Kennard–Stone (KS) algorithm, which describes the differences among samples with Euclidean distance, is suitable for complex and multidimensional data in qualitative analysis (Galvão *et al.*, 2005). In this study, the samples were divided into calibration set and prediction set according to a ratio of 3:1 using the KS algorithm. To ensure that the samples at different mastitis degrees were distributed evenly, the division was separately conducted on the samples at each mastitis degree.

Variable selection

Variable selection is usually used to remove redundant variables and keep useful information as much as possible. In this study, the variable importance in the projection (VIP) method based on partial least squares (PLS) regression was used to evaluate the contribution of each x variable (i.e., spectral data here) to the y variance (i.e., mastitis degree here). The x variable with a small VIP value has little ability to interpret the variance of y. The redundant variables at different levels in spectra can be effectively identified with different thresholds.

Data reduction

Principal component analysis (PCA) and PLS are widely used dimension reduction methods that can effectively process the data with a large amount of correlation or co-linearity (Han *et al.*, 2022). In these methods, the feature space is substituted with the relatively low dimension projected space, called principal components (PCs) and latent variables (LVs), respectively. This process can reduce noises and redundant information with minimal loss of useful information and improve model generalization ability (Liu and Guo, 2014). The PCs obtained using PCA are ordered according to the decreased explained variance in a spectral matrix. However, LVs in PLS are ordered according to their relevance in predicting the y variables. Compared with the first several PCs, the first several LVs may be more informative concerning the response variable (Nicolaï *et al.*, 2007). In this study, the PCs and LVs obtained using PCA and PLS, respectively, were used as the input of models to simplify models and improve the prediction accuracy of models.

Modeling methods

In this study, linear discriminant analysis (LDA), support vector classification (SVC), and feed-forward neural network (FFNN) were used to establish qualitative analysis models for identifying mastitis degrees.

LDA

LDA is a linear classification model that supports multiclass in principle. It aims to build linear combinations of a feature set to characterize the interests of different classes. The classification is achieved with a linear classifier based on the feature combinations. Compared with complex classification methods, LDA is simpler, more robust, and might provide similar model accuracy (Naderi-Boldaji *et al.*, 2018).

SVC

SVC is a highly robust and efficient algorithm for two-class classification, which establishes a decision boundary in the feature space for separating data points into different classes. For non-separable problems, various kernel functions are designed to map the original feature space into high-dimensional feature space for obtaining a clear separation between different classes. Then, SVC could search a hyperplane in the high-dimensional feature space, maximizing the margin and minimizing the number of misclassified samples (Raghavendra and Deka, 2014). The radial basis function, which can simplify the training procedure and provide better performance than other functions (Pang *et al.*, 2022), was selected as the kernel function in this study. For an n-class classification, $n\times(n\text{-}1)$ SVC models are built, and each built model is only used to distinguish two classes. When making a prediction, the class with the most votes is the predicted results of the sample.

FFNN

As a commonly used artificial neural network, FFNN is built with neurons organized in several layers, where the neurons in each layer are fully connected to the next layer. The numbers of neurons in the input and output layers are determined by the dimensions of input and output variables, respectively. A non-linear system might be accurately approximated with an FFNN containing one hidden layer with a non-linear activation function (González-Viveros *et al.*, 2021). Therefore, an FFNN with one hidden layer was built in this study. The number of neurons in the hidden layer was determined by

Table 1. Statistical results on SCC, main components, pH, and electrical conductivity (mean±SD) of used raw milk samples in this study

Mastitis grade	No. of samples	SCC[1] (×10³ cells/mL)			Fat (%)	Protein (%)	Lactose (%)	TS (%)	pH	EC (mS/cm)
		Min.	Max.	Mean±SD						
Negative	113	5	494	141±145.05[c]	3.43±1.06[a]	3.16±0.46[a]	5.06±0.16[a]	11.97±1.5[a]	6.61±0.75[a]	3.83±0.74[b]
Weakly positive	92	509	1435	807±227.66[b]	3.23±0.74[a]	3.06±0.30[b]	4.88±0.14[b]	11.56±0.99[a]	6.61±0.80[a]	4.11±0.58[b]
Positive	96	1514	4971	2553±865.19[a]	3.31±0.53[a]	3.18±0.37[a]	4.74±0.24[c]	11.72±0.69[a]	6.69±0.11[a]	4.65±0.55[a]

The different character in the same column means a significant difference at the significance level of 0.05.
[1] The mean is geometric mean.
SD: standard deviation; TS: total solids; EC: electrical conductivity.

the highest accuracy rate of the built FFNN models for the prediction set. Softmax activation function and cross-entropy loss function were used in the output layer, whereas a rectifier linear unit activation function was applied in all other layers.

Model performance assessment

The recall and precision rates of each mastitis degree, calculated using Eq. 1 and Eq. 2, respectively, were used to evaluate the prediction performance of established models for each degree. When calculating the recall and precision rates of each degree in multiclass classification models, the category of each sample was converted into the target category or other categories.

$$\text{Recall rate} = \text{TP}/(\text{TP} + \text{FN}) \quad (1)$$

$$\text{Precision rate} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

where TP, true positives, and FN, false negatives, represent that a sample in the target category was identified as a target category and other categories, respectively. TN, true negative, and FP, false positives, indicate that a sample in other categories was classified as other categories and target category, respectively. The accuracy rate, the ratio of correctly predicted samples to total samples, was used to evaluate the performance of classification models for calibration and prediction sets.

Software

Besides the dielectric spectra collection software 85070D mentioned above, SPSS 17.0 (SPSS Inc., Chicago, IL, USA) was used for one-way analysis of variance (ANOVA) on the SCC, main compositions, pH, electrical conductivity, and permittivities (ε' and ε'') of milk samples using Tamhane T2 test at $p<0.05$ to investigate the significant difference among different mastitis degrees. Except for that noted specially, the significance level in this research refers to 0.05. Spyder 3.2.6 matched with Python 3.8 and Scikit-Learn (Pedregosa *et al.*, 2011) was applied for spectra preprocessing and analysis, variable selection, data reduction, and establishment of LDA and SVC models. Keras (Chollet, 2015), a deep learning library, was used to construct and optimize FFNN networks.

Results and discussion

Statistical results of used milk samples

According to the obtained SCC of each milk sample and the Chinese industry standard on subclinical mastitis in dairy cow (MARA, 2015), there were 113, 92, and 96 samples at negative, weakly positive, and positive mastitis degrees, respectively. Table 1 shows the statistical results on SCC, main compositions, pH, and electrical conductivity of the 301 milk samples used. The SCC of all milk samples was between 5×10^3 cells/mL and 4971×10^3 cells/mL. The geometric mean SCCs of negative, weakly positive, and positive samples were 141×10^3, 807×10^3, and 2553×10^3 cells/mL, respectively. The SCC of the milk at three mastitis degrees showed a significant difference.

The fat, total solids, and pH showed no significant difference among the three mastitis degrees. Sunds *et al.* (2021) reported that mastitis causes a decrease in fat and total solids as a result of reduced secretory capacity. However, mastitis also reduces milk production, which could increase the concentration of each component (Goncalves *et al.*, 2020). Therefore, there were no differences in fat and total solids among the three mastitis degrees in this study. A similar result of pH was found in another study (Ogola *et al.*, 2007). The protein content of weakly positive samples (3.06%±0.30%) was significantly lower than that of the negative (3.16%±0.46%) and positive (3.18%±0.37%) samples. The negative samples had the highest lactose content (5.06%±0.16%) and the positive samples had the lowest lactose content (4.74%±0.24%). Moreover, a significant difference was noted for the lactose content of three mastitis degrees. The electrical conductivity of the positive samples (4.65±0.55 mS/cm) was significantly higher than that of the weakly positive (4.11±0.58 mS/cm) and negative (3.83±0.74 mS/cm) samples. The results showed that the electrical conductivity of milk is an effective index in distinguishing positive mastitis from negative and weakly positive mastitis and indicated that lactose content has the potential in classifying mastitis degrees. Norberg *et al.* (2004) used the electrical conductivity to predict mastitis degrees. The result indicated that the accuracy of clinical mastitis (80.6%) was obviously higher than that of subclinical mastitis (45.0%) and agreed with the result of this study.

Effect of SCC on dielectric properties

The mean ε' and ε'' spectra of milk samples at three mastitis degrees over the frequency range of 20–4500 MHz are shown in Figure 1. The ε' of all samples at different mastitis degrees decreased with increasing frequency over the whole frequency range, and the decrease was quick at frequencies below approximately 40 MHz (Figure 1A). The ε' decreased first below about 1800 MHz and then increased slightly above 1800 MHz with increasing frequency (Figure 1B). The decreased ε' resulted from the dipole motion, which

could not follow the changed direction of the electric field when the frequency increased (El Khaled *et al.*, 2016). Ionic conduction and dipole relaxation are the main dielectric losses in the low- and high-frequency ranges, respectively (Ohlsson, 1989). The former caused ε'' to decrease rapidly with the increase of frequency in the low-frequency range, while the latter led to an increase in ε'' with increasing frequency above 1800 MHz. Similar phenomena were reported for milk in other studies (Guo *et al.*, 2017; Zhu *et al.*, 2018).

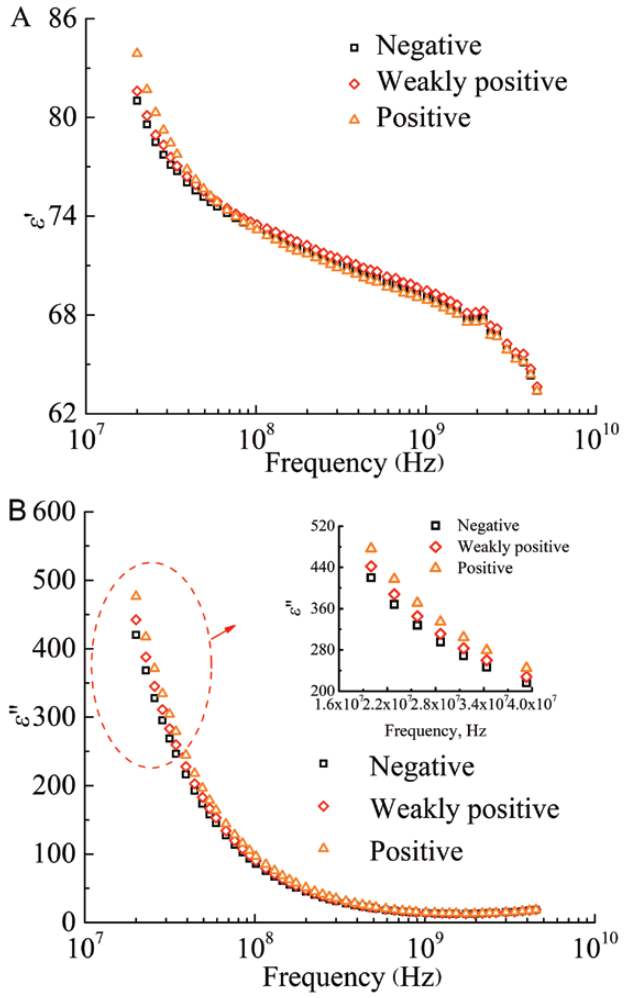

Figure 1. The mean dielectric spectra of ε′ (A) and ε″ (B) of the milk samples at three mastitis degrees.

Although the mastitis degrees or SCC of milk samples did not affect the changing trend of permittivity with frequency, the negative samples had the lowest and the positive samples had the highest ε' and ε'' below approximately 50 MHz. Over the whole frequency range, the positive samples had the largest decrease, while the negative samples had the smallest decrease in ε'. For example, ε' and ε'' at 27.12 MHz were 78.07±1.79 and 310.81±24.90 for negative samples, 78.49±1.06 and 327.41±14.25 for weakly positive samples, and 79.63±1.46 and 351.94±27.04 for positive samples. When the frequency increased from 20 MHz to 4500 MHz, the ε' of negative samples decreased from 81.02±2.12 to 63.38±1.87, while the ε' of positive samples decreased from 83.88±2.28 to 63.38±1.29.

Table 2 lists the mean values and standard deviations of ε' and ε'' of milk samples at three mastitis degrees and six selected frequencies, including 27.12, 40.68, 100, 500, 915, and 2450 MHz. The ANOVA results showed that at 27.12 MHz, the ε' values of the positive samples were significantly different from those of the weakly positive and negative samples, while at 40.68 MHz, the ε' value of the positive samples was significantly different from that of the negative samples. At 100 MHz, the ε' of the milk samples at the three mastitis degrees had no significant difference, but the ε' of positive samples had a significant difference with that of the weakly positive samples at 500, 915, and 2450 MHz. Except for 2450 MHz, the ε' of the milk samples at three mastitis degrees had significant differences. At 2450 MHz, the ε'' value of the weakly positive samples was significantly different from those of the negative and positive samples.

Milk can be considered an aqueous solution with somatic cells in suspension. With the increase of SCC, ionic concentration induced by inflammation increased, which increased ε' in the low-frequency range (Grillo *et al.*, 2002). Meanwhile, mastitis results in increasing polar molecules, such as whey protein, free fatty acid, sodium, and chloride, due to mixing with blood components (Ogola *et al.*, 2007), which play a decisive role in dielectric loss. Therefore, compared with negative mastitis samples, the positive or weakly positive mastitis samples have larger ε''.

Spectra preprocessing and analysis

The original dielectric spectra, including 402 dielectric variables (201 variables of ε' and 201 variables of ε''), were preprocessed using SG smoothing, SNV, and SG+SNV. After

Table 2. Dielectric constants and loss factors (mean±SD) of raw milk samples at six selected frequencies and three mastitis degrees

Frequency (MHz)	ε'			ε''		
	Negative	Weakly positive	Positive	Negative	Weakly positive	Positive
27.12	78.07±1.79^{b}	78.49±1.06^{b}	79.63±1.46^{a}	310.81±24.90^{c}	327.41±14.25^{b}	351.94±27.04^{a}
40.68	75.93±1.90^{b}	76.29±1.35a^{b}	76.70±0.97^{a}	209.78±16.85^{c}	220.89±9.66^{b}	237.31±18.17^{a}
100	73.25±2.11^{a}	73.48±1.69^{a}	73.16±1.19^{a}	85.73±6.83^{c}	90.23±3.91^{b}	96.75±7.24^{a}
500	70.33±2.25^{ab}	70.68±1.86^{a}	70.03±1.29^{b}	20.68±1.43^{c}	21.58±0.80^{b}	22.80±1.54^{a}
915	69.28±2.23^{ab}	69.62±1.85^{a}	69.02±1.28^{b}	14.48±0.82^{c}	14.99±0.45^{b}	15.54±0.88^{a}
2450	66.83±2.24^{ab}	67.28±1.80^{a}	66.74±1.15^{b}	13.83±0.72^{b}	14.15±0.58^{a}	13.97±0.38^{b}

The different character in the same row for dielectric constant and loss factor mean a significant difference at the significance level of 0.05.
SD: standard deviation.

Table 3. The determined number of latent variables and obtained accuracy rates of built LDA models with PLS as data reduction method at different spectra preprocessing and outlier detection methods

Preprocessing method	Outlier detection method	Number of latent variables	Accuracy rate (%)	
			Calibration set	Prediction set
SG	LOF	20	82.8	87.3
	1SVM	20	82.8	85.9
SNV	LOF	20	83.6	90.3
	1SVM	23	86.0	88.7
SG+SNV	LOF	25	87.4	93.1
	1SVM	27	90.2	88.7

preprocessing, the outlier detections were conducted based on LOF and 1SVM. The probability of finding a new observation outside the frontier for both outlier detection methods was set as 5%. The neighbor number in LOF was determined to be 20, and the linear kernel was chosen as the kernel function of 1SVM. Finally, 16 different outliers were identified by using LOF and 1SVM, respectively.

The remaining 286 samples were divided into calibration and prediction sets according to a ratio of 3:1 using the KS algorithm at each mastitis degree. LDA models with PLS as a data reduction method were built with different preprocessing and outlier detection methods. The inputs of LDA models were preprocessed dielectric spectra, and their outputs were dummy variables, i.e., 0, 1, and 2, which were assigned artificially to represent negative, weakly positive, and positive samples, respectively. The number of LVs used for the LDA model obtained using PLS was determined by cross-validation based on the lowest value of the root-mean-squares error of cross-validation (RMSECV).

Table 3 lists the identification accuracy rates of the established LDA models for samples in the calibration and prediction sets. This indicates that the LDA model built with the preprocessed spectra using SG+SNV and LOF as an outlier detection method had the best accuracy rate of 93.1% for the prediction set. Therefore, SG+SNV and LOF were used as the dielectric spectra preprocessing and outlier detection methods, respectively, in further analysis. Finally, the calibration set had 214 samples (77 negative samples, 68 weakly positive samples, and 69 positive samples), and the prediction set had 72 samples (26 negative samples, 23 weakly positive samples, and 23 positive samples).

Variable selection

The PLS regression models at different numbers of LVs from 1 to 35 with an interval of 1 were established based on 214 milk samples in the calibration set by applying the cross-validation method. The PLS regression model with 25 LVs, which had the lowest RMSECV, was chosen to calculate the VIP scores. The results showed that the VIP scores of 402 dielectric variables were between 0.76 and 2.40. When the thresholds were set as 0.7, 0.8, 0.9, and 1.0, the determined numbers of variables were 402 (all variables), 372, 241, and 131, respectively. The accuracy rates of the built LDA models based on these selected variables were 93.1%, 93.1%, 91.7%, and 91.7%, respectively, for the prediction set. Although the accuracy rates were the same (93.1%) when the threshold was 0.7 and 0.8, the number of variables was less when the threshold was 0.8. Finally, 0.8 was set as the threshold of VIP, and the dielectric spectra with 372 effective variables, including 182 variables of ε' and 190 variables of ε'', were applied for subsequent data reduction.

Data reduction

Figures 2A and 2B show the first two PCA scores (PC1 and PC2) with an accumulative contribution rate of 99.74% and the values of the first two LVs (LV1 and LV2) in PLS for the milk samples in the calibration set, respectively. The results showed that the milk samples at three mastitis degrees mixed considerably. It is not easy to separate the three degree samples based on the first two PCs or LVs. Figure 3 shows the accuracy rates of established LDA models for the samples in the calibration set using different numbers of PCs and LVs. It shows that when the numbers of PCs and LVs reached 22 and 24, the accuracy rates were highest, with 79.4% and 86.9%, respectively. Therefore, 22 PCs obtained by PCA and 24 LVs obtained by PLS were used to build different qualitative models.

Modeling results

The LDA, SVC, and FFNN models were built for identifying the samples at three mastitis degrees when the 22 PCs calculated by PCA and 24 LVs obtained by PLS were used as inputs of these models. The dummy variables, 0, 1, and 2, standing for negative, weakly positive, and positive samples, respectively, were used as the output of the LDA and SVC models. For the FFNN model, the dummy variables were converted to a Boolean matrix with three columns in which only the column of the corresponding label appeared was 1, and the rest was 0.

LDA modeling results

Table 4 shows the classification results of the built LDA models when PCA and PLS were used as data reduction methods. The results showed that the recall rates of the established LDA model based on PLS (PLS-LDA) were 96.2%, 87.0%, and 100%, and the precision rates were 92.6%, 95.2%, and 95.8% for negative, weakly positive, and positive samples in the prediction set, respectively. The PLS-LDA's recall and precision rates were higher than or equal to those of the LDA model based on PCA (PCA-LDA). Finally, the accuracy rates of PLS-LDA for the calibration and prediction

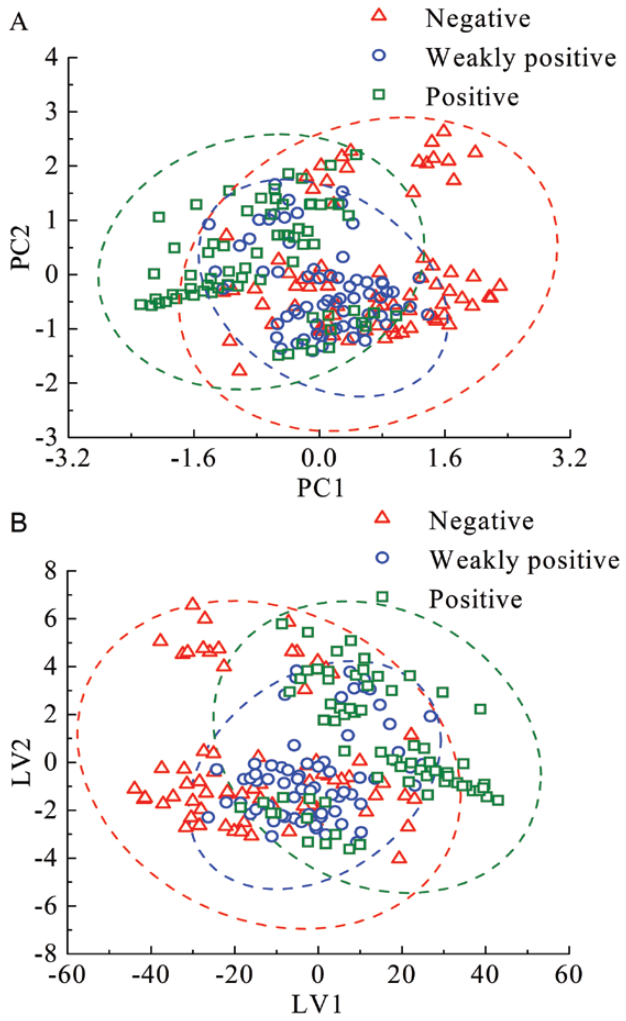

Figure 2. The principal components (PCs) (A) and latent variables (LVs) (B) plots of the milk samples at three mastitis degrees in the calibration set.

sets (86.9% and 94.4%, respectively) were higher than those of the PCA-LDA model (79.4% and 93.1%, respectively). It was noted that the recall of the built LDA models for positive samples was 100%, and the identification performance for weakly positive samples was poorer than the for negative and positive samples.

SVC modeling results

The penalty factor *c* and slack variable *g* are essential parameters of the SVC model with the radial basis kernel function. The optimal *c* and *g* were selected from 2^{-10} to 2^{10} with an interval of $2^{0.2}$ using grid search and cross-validation methods. The best *c* and *g* were determined by the highest accuracy rate of built SVC models. In this study, the optimal *c* and *g* were 84.45 and 0.0039 for the SVC model based on PCA (PCA-SVC) and 388.02 and 0.0034 for the SVC model based on PLS (PLS-SVC), respectively.

The classification results of the established SVC models for the samples at different mastitis degrees are listed in Table 4. For the samples in the prediction set, the recall rates of PCA-SVC were 92.3%, 95.7%, and 100%, and the precision rates were 96.0%, 91.7%, and 100%, for the negative, weakly positive, and positive samples, respectively. The accuracy

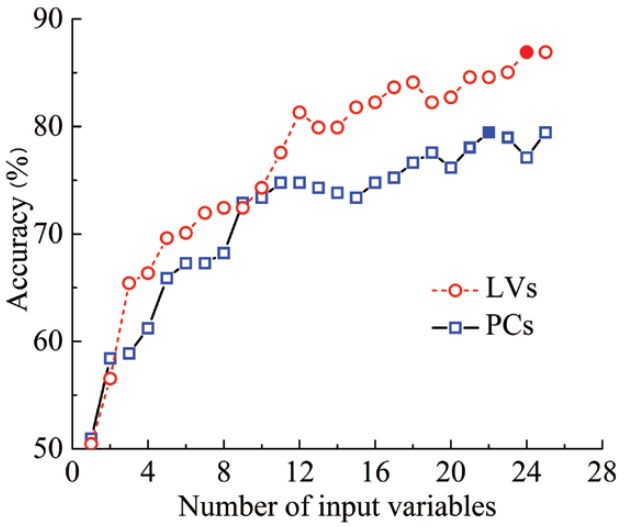

Figure 3. The accuracy rates of established linear discriminant analysis (LDA) models for the samples in the calibration set at different numbers of latent variables (LVs) and principal components (PCs).

rates of PCA-SVC for calibration (91.6%) and prediction sets (95.8%) were higher than 88.3% and 94.4% of PLS-SVC, respectively.

FFNN modeling results

Table 4 shows the classification results of built FFNN models with 80 and 40 neurons in the hidden layers of the built FFNN models based on PCA (PCA-FFNN) and PLS (PLS-FFNN), respectively. The results showed that the recall rates of PCA-FFNN for the prediction set were 96.2%, 91.3%, and 95.7%, and the precision rates were 92.6%, 91.3%, and 100% for the negative, weakly positive, and positive samples, respectively. The accuracy rate of PCA-FFNN (94.4%) for the prediction set was higher than that of PLS-FFNN (91.7%).

Comparison

When the two data reduction methods were compared, it was noted that PLS performed better than PCA for the LDA models. However, PCA performed better than PLS for the SVC and FFNN models. This indicated that the PLS based on correlation ranking was more suitable for simple linear classification models and can provide the same predictive performance as complex non-linear models. Among the six established models, PCA-SVC had the best prediction performance with an accuracy rate of 95.8%, followed by PLS-LDA, PLS-SVC, and PC-FFNN with 94.4%.

In contrast to the reported sensitivities of 80.6% for clinical mastitis cases and 45.0% for subclinical mastitis cases, and also with the precision rate of 74.8% for healthy cases using an electrical conductivity index (Norberg *et al.*, 2004), the reported sensitivities (recall) here for positive (100%) and weakly positive (95.7%) samples and the precision (96.0%) for negative samples obtained by the PCA-SVC model were higher. The obtained sensitivity, specificity, and accuracy rates in this study were also higher than the reported data ranging in 68%–84%, 60%–85%, and 56%–81%, respectively, when the six best electrical conductivity indexes were used to identify clinical mastitis (Khatun *et al.*, 2017). In contrasted to the obtained 77.78% sensitivity

Table 4. Classification results of the established LDA, SVC, and FFNN models using PCA and PLS as data reduction methods

Model	Data reduction method	Calibration set							Prediction set						
		N		WP		P		AC	N		WP		P		AC
		RE	PR	RE	PR	RE	PR		RE	PR	RE	PR	RE	PR	
LDA	PCA	67.5	85.2	77.9	74.6	94.2	79.3	79.4	92.3	92.3	87.0	90.9	100	95.8	93.1
	PLS	81.8	91.3	80.9	82.1	98.6	87.2	86.9	96.2	92.6	87.0	95.2	100	95.8	94.4
SVC	PCA	88.3	94.4	88.2	89.6	98.6	90.7	91.6	92.3	96.0	95.7	91.7	100	100	95.8
	PLS	85.7	88.0	80.9	85.9	98.6	90.7	88.3	96.2	96.2	91.3	91.3	95.7	95.7	94.4
FFNN	PCA	93.5	93.5	92.6	94.0	98.6	97.1	94.9	96.2	92.6	91.3	91.3	95.7	100	94.4
	PLS	92.2	94.7	89.7	93.8	100	93.2	93.9	96.2	89.3	82.6	90.4	95.7	95.7	91.7

AC: accuracy rate; N: negative; P: positive; PR: precision rate; RE: recall rate; WP: weakly positive.

and 80.56% specificity reported by Meilina *et al.* (2009) for distinguishing normal milk and mastitis milk using near-infrared spectroscopy, the reported sensitivity and specificity here were also higher. In general, the milk SCC measured by SCC counters dependent on flow cytometry, in which there is no qualitative analysis model. Based on these SCC results, early warning of clinical and subclinical mastitis could be provided after complicated analysis. However, the PCA-SVC model with dielectric spectra could directly output degree of mastitis in this study. Therefore, this indicated that the dielectric spectra of raw milk have great potential in the rapid detection of mastitis degrees of dairy cows.

Conclusions

For rapid identification of the mastitis degrees of cows by raw milk, dielectric spectra and SCC of 301 raw milk samples were measured using a dielectric measurement system based on coaxial probe technology and a milk component analyzer. SG+SNV was used for preprocessing dielectric spectra, and outlier detection was conducted using LOF. The remaining 286 samples were divided into calibration and prediction sets using the KS algorithm at a ratio of 3:1. A total of 372 variables in dielectric spectra, including 182 variables of ε' and 190 variables of ε'', were selected by VIP values for data dimension reduction. After dimension reduction with PCA and PLS, the LDA, SVC, and FFNN models were built to identify the milk samples at three mastitis degrees, including negative, weakly positive, and positive. The results indicated that the PCA+SVC model had the best prediction ability, with an accuracy rate of 95.8% for the prediction set. The models of PLS-LDA, PLS-SVC, and PCA-FFNN also had excellent prediction ability, with an accuracy rate of 94.4% for the prediction set. The study shows that the dielectric spectra of raw milk have great potential in identifying negative, weakly positive, and positive mastitis of cows, and it provides a feasible, quick, and *in situ* mastitis detection method for farms to monitor the health status of cows. Dielectric spectra are greatly affected by sample temperature. The dielectric spectra of raw milk in this study were obtained at 25 °C, which is much lower than that of raw milk in-line. The effect of temperature should be considered in practical mastitis identification using dielectric spectra. It should also be noted that dielectric spectra could only identify cows with mastitis when the milk SCC is higher than 500×10^3 cells/mL. However, glandular infection may have occurred at around 200 000 cells/mL.

Author Contributions

Zhuozhuo Zhu: Methodology, formal analysis, validation, writing original draft, review and editing. Biying Lin: Methodology, formal analysis, validation, writing original draft, review and editing. Xinhua Zhu: Resources, project administration, funding acquisition, writing, review and editing. Wenchuan Guo: Resources, data curation, validation, supervision, writing, review and editing.

Funding

The study was financially supported by the National Natural Science Foundation of China (No. 32172308 and No. 31671935).

Conflict of Interest

Zhuozhuo Zhu, Biying Lin, Xinhua Zhu, and Wenchuan Guo declare that they have no conflict of interest.

References

Afseth, N. K., Segtnan, V. H., Wold, J. P. (2006). Raman spectra of biological samples: a study of preprocessing methods. *Applied Spectroscopy*, 60(12): 1358–1367.

Altemimi, A., Aziz, S. N., Al-HiIphy, A. R. S., *et al.* (2019). Critical review of radio-frequency (RF) heating applications in food processing. *Food Quality and Safety*, 3(2): 81–91.

Atasever, S., Erdem, H., Kul, E. (2012). Using viscosity values for determining somatic cell count in cow milk. *Asian Journal of Animal and Veterinary Advances*, 7(5): 441–445.

Becheva, Z. R., Gabrovska, K. I., Godjevargova, T. I. (2018). Comparison between direct and indirect immunofluorescence method for determination of somatic cell count. *Chemical Papers*, 72(8): 1861–1867.

Breunig, M., Kriegel, H. P., Ng, R., *et al.* (2000). LOF: identifying density-based local outliers. The 2000 ACM SIGMOD International Conference on Management of Data, 15–18 May 2000, Dallas, TX, USA, pp. 93–104.

Chollet, F. (2015), Keras, GitHub repository [Online]. https://github.com/fchollet/keras. Accessed on 12 November 2022.

El Khaled, D., Castellano, N. N., Gazquez, J. A., *et al.* (2016). Dielectric spectroscopy in biomaterials: agrophysics. *Materials*, 9(5): 310.

Frundzhyan, V. G., Parkhomenko, I. M., Brovko, L. Y., *et al.* (2008). Improved bioluminescent assay of somatic cell counts in raw milk. *Journal of Dairy Research*, 75(3): 279–283.

Galvão, R. K. H., Araujo, M. C. U., José, G. E., *et al.* (2005). A method for calibration and validation subset partitioning. *Talanta*, 67(4): 736–740.

Goncalves, J. L., Kamphuis, C., Vernooij, H., *et al.* (2020). Pathogen effects on milk yield and composition in chronic subclinical mastitis in dairy cows. *Veterinary Journal*, 262: 105473.

González-Viveros, N., Gómez-Gil, P., Castro-Ramos, J., *et al.* (2021). On the estimation of sugars concentrations using Raman spectroscopy and artificial neural networks. *Food Chemistry*, 352: 129375.

Grillo, G. J., Perez, M. A., Anton, J. C., *et al.* (2002). Direct-evaluation of the fresh-milk somatic cell concentration (SCC) through electrical permittivity measurements. 19th IEEE Instrumentation and Measurement Technology Conference. 21–23 May 2002, Anchorage, AK, USA, pp. 975–979.

Guo, W., Fang, L., Liu, D., *et al.* (2015). Determination of soluble solids content and firmness of pears during ripening by using dielectric spectroscopy. *Computers and Electronics in Agriculture*, 117: 226–233.

Guo, W., Lin, B., Liu, D., *et al.* (2017). A novel technique on determining water content in milk using radio-frequency/microwave dielectric spectroscopy and chemometrics. *Food Analytical Methods*, 10: 3781–3789.

Guo, Z., Wang, M., Agyekum, A. A., *et al.* (2020). Quantitative detection of apple watercore and soluble solids content by near infrared transmittance spectroscopy. *Journal of Food Engineering*, 279: 109955.

Han, Z., Li, B., Wang, Q., *et al.* (2022). Detection of skin defects on loquat using the hyperspectral imaging combining both band radio and improved three-phase level set segmentation method. *Food Quality and Safety*, 7: fyac065.

IDF (International Diabetes Federation). (2013). Guidelines for the use and interpretation of bovine milk somatic cell counts (SCC) in the dairy industry. *Bulletin of the International Dairy Federation*, 2013(466): 1–15.

Iweka, P., Kawamura, S., Mitani, T., *et al.* (2020). Online milk quality assessment during milking using near-infrared spectroscopic sensing system. *Environmental Control in Biology*, 58(1): 1–6.

Jafari, S., Dehghani, M., Nasirizadeh, N., *et al.* (2019). Synthesis and characterisation of a selective adsorbent based on the molecularly imprinted polymer for the removal of cloxacillin antibiotic residue from milk. *International Journal of Dairy Technology*, 72(4): 505–514.

Khatun, M., Clark, C. E. F., Lyons, N. A., *et al.* (2017). Early detection of clinical mastitis from electrical conductivity data in an automatic milking system. *Animal Production Science*, 57(7): 1226–1232.

Lien, C. C., Wan, Y. N., Ting, C. H. (2016). Online detection of dairy cow subclinical mastitis using electrical conductivity indices of milk. *Engineering in Agriculture, Environment and Food*, 9(3): 201–207.

Liu, D., Guo, W. (2014). Identification of kiwifruits treated with exogenous plant growth regulator using near-infrared hyperspectral reflectance imaging. *Food Analytical Methods*, 8(1): 164–172.

MARA (Ministry of Agriculture and Rural Affairs of the People's Republic of China). (2004). Enumeration of somatic cells in raw milk, NY/T800-2004. MARA, Beijing, China.

MARA (Ministry of Agriculture and Rural Affairs of the People's Republic of China). (2015). Rapid diagnostic techniques for subclinical mastitis in dairy cow, NY/T2692-2015. MARA, Beijing, China.

Mariani, E., Cipolat-Gotet, C., Stefanon, B., *et al.* (2022). Effect of total and differential somatic cell count on yield, composition and predicted coagulation properties from individual dairy cows. *International Journal of Dairy Technology*, 75(2): 298–307.

Meilina, H., Kuroki, S., Jinendra, B. M., *et al.* (2009). Double threshold method for mastitis diagnosis based on NIR spectra of raw milk and chemometrics. *Biosystems Engineering*, 104(2): 243–249.

Melfsen, A., Hartung, E., Haeussermann, A. (2012). Accuracy of in-line milk composition analysis with diffuse reflectance near-infrared spectroscopy. *Journal of Dairy Science*, 95(11): 6465–6476.

Naderi-Boldaji, M., Mishra, P., Ahmadpour-Samani, M., *et al.* (2018). Potential of two dielectric spectroscopy techniques and chemometric analyses for detection of adulteration in grape syrup. *Measurement*, 127: 518–524.

Naito, H., Ogawa, Y., Kubota, A., *et al.* (2013). Measurement of somatic cell count in cow's raw milk for detecting mastitis by using attenuated total reflection terahertz spectroscopy. *IFAC Proceedings Volumes*, 46(18): 331–335.

Naito, H., Ogawa, Y., Kubota, A., *et al.* (2015). Attenuated total reflectance terahertz spectroscopy for quantitative measurement of somatic cell count in bovine milk. *Engineering in Agriculture, Environment and Food*, 8(2): 79–82.

Nicolaï, B. M., Beullens, K., Bobelyn, E., *et al.* (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review. *Postharvest Biology and Technology*, 46(2): 99–118.

Norberg, E., Hogeveen, H., Korsgaard, I. R., *et al.* (2004). Electrical conductivity of milk: ability to predict mastitis status. *Journal of Dairy Science*, 87(4): 1099–1107.

Ogola, H., Shitandi, A., Nanua, J. (2007). Effect of mastitis on raw milk compositional quality. *Journal of Veterinary Science*, 8(3): 237–242.

Ohlsson, T. (1989). Dielectric properties and microwave processing. In: Singh R. P, Medina A. G. (Eds.). *Food Properties and Computer-Aided Engineering of Food Processing Systems*. Springer, Dordrecht, The Nethelands, pp. 73–92.

Pang, L., Chen, H., Yin, L., *et al.* (2022). Rapid fatty acids detection of vegetable oils by Raman spectroscopy based on competitive adaptive reweighted sampling coupled with support vector regression. *Food Quality and Safety*, 6: fyac053.

Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Raghavendra, N. S., Deka, P. C. (2014). Support vector machine applications in the field of hydrology: a review. *Applied Soft Computing*, 19: 372–386.

Rinnan, A., Berg, F., Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10): 1201–1222.

Sanaeifar, A., Jafari, A., Golmakani, M. T. (2018). Fusion of dielectric spectroscopy and computer vision for quality characterization of olive oil during storage. *Computers and Electronics in Agriculture*, 145: 142–152.

Savitzky, A., Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8): 1627–1639.

Schalm, O. W., Noorlander, D. O. (1957). Experiments and observations leading to development of the California mastitis test. *Journal of the American Veterinary Medical Association*, 130(5): 199–204.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., *et al.* (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7): 1443–1471.

Sunds, A. V., Sundekilde, U. K., Poulsen, N. A., *et al.* (2021). Association between udder inflammation and glycosidase activities and free sugar levels in bovine milk. *International Dairy Journal*, 120: 105093.

Thennadil, S. N., Dewar, M., Herdsman, C., *et al.* (2018). Automated weighted outlier detection technique for multivariate data. *Control Engineering Practice*, 70: 40–49.

Thiruvengadam, M., Venkidasamy, B., Selvaraj, D., *et al.* (2020). Sensitive screen-printed electrodes with the colorimetric zone for simultaneous determination of mastitis and ketosis in bovine milk samples. *Journal of Photochemistry and Photobiology B: Biology*, 203: 111746.

Tsenkova, R., Atanassova, S., Kawano, S., *et al.* (2001). Somatic cell count determination in cow's milk by near-infrared spectroscopy: a new diagnostic tool. *Journal of Animal Science*, 79(10): 2550–2557.

Wickström, E., Persson-Waller, K., Lindmark-Månsson, H., *et al.* (2009). Relationship between somatic cell count, polymorphonuclear leucocyte count and quality parameters in bovine bulk tank milk. *Journal of Dairy Research*, 76(2): 195–201.

Zhang, S., Zhou, L., Ling, B., *et al.* (2016). Dielectric properties of peanut kernels associated with microwave and radio frequency drying. *Biosystems Engineering*, 145: 108–117.

Zhu, X., Guo, W., Liu, D., *et al.* (2018). Determining the fat concentration of fresh raw cow milk using dielectric spectroscopy combined with chemometrics. *Food Analytical Methods*, 11(5): 1528–1537.

Zhu, Z., Zhu, X., Kong, F., *et al.* (2019). Quantitatively determining the total bacterial count of raw goat milk using dielectric spectra. *Journal of Dairy Science*, 102(9): 7895–7903.