September 2024 Vol. 31 No. 3 59–73
Journal of Northeast Agricultural University (English Edition)

Rapid Detection of Somatic Cell Count Based on Hybrid Variable Selection Method

Shen Weizheng[1], Cui Xiang[1], Wang Yan[1*], Nie Debao[2], Zhang Qinggang[2], Zheng Wei[2], Sun Jian[1, 3], Yang Xin[1], and Dai Baisheng[1]

[1] College of Electrical and Information, Northeast Agricultural University, Harbin 150030, China

[2] Heilongjiang Animal Husbandry Service, Harbin 150060, China

[3] School of Information Engineering, Shandong Huayu University of Technology, Dezhou 253000, Shandong, China

Abstract: Somatic cell count detection is the daily work of dairy farms to monitor the health of cows. The feasibility of applying near-infrared spectroscopy to somatic cell count detection was researched in this paper. Milk samples with different somatic cell counts were collected and preprocessing methods were studied. Variable selection algorithm based on hybrid strategy and modelling method based on ensemble learning were explored for somatic cell count detection. Detection model was used to diagnose subclinical mastitis and the results showed that near-infrared spectroscopy could be a tool to realize rapid detection of somatic cell count in milk.

Key words: near-infrared spectroscopy, somatic cell count, mastitis, rapid detection

CLC number: S156.2 **Document code:** A **Article ID:** 1006–8104(2024)–03–0059–15

Introduction

Somatic cell count (SCC) detection in raw milk is an important daily work of dairy farms. It is a major means to monitor the health of cows, especially the occurrence of mastitis. Mastitis is one of the most common diseases of dairy cows, its appearance will significantly reduce the milk production and quality of dairy cows, resulting in an increase in costs and a decrease in profit (Halasa *et al*., 2007; He *et al*., 2020; Albenzio *et al*., 2011). SCC refers to the total number of somatic cells per milliliter of raw milk and it should be tested regularly to monitor the health of cows in order to diagnose mastitis as early as possible, especially subclinical mastitis. SCC is also closely related to the quality of milk (Dos Reis *et al*., 2011; Ma *et al*., 2000). The large number of samples, high cost and accuracy of results are problems for SCC detection in dairy farms.

Many scholars at home and abroad are committed to SCC detection. At present, many detection methods have emerged, including the direct detection method and the indirect detection method. Direct detection refers to counting milk somatic cells using different techniques and methods, including manual microscopy, flow cytometry and computer vision technology (Keisler *et al*., 1992; Zhang *et al*., 2018). However, the expensive price of the instrument makes it impossible to be widely promoted and used in dairy farms. Indirect detection methods are according to changes of

Supported by the Natural Science Foundation of Heilongjiang Province of China (LH2023C016); the Key Research and Development Program of Heilongjiang Province of China (2022ZX01A24); the National Modern Agricultural Industry Technology System (CARS36)
Shen Weizheng (1977–), male, Ph. D, professor, engaged in the research of intelligent animal husbandry. E-mail:wzshen@neau.edu.cn
* Corresponding author. E-mail: wangyan_neau@126.com

composition, conductivity and pH in milk with SCC increasing. Researchers estimate SCC by establishing a correlation between SCC and these changes, such as the California Mastitis Test Reagent (CMT), Wisconsin mastitis test method and DNA method (Ashraf *et al*., 2018; Urbanová *et al*., 1985). Preprocessing of samples is usually required and results in complex operation and longer detection time. Low cost is the advantage of the above methods, but heavy workload and subjective evaluation are major issues resulting in these methods unsuitable for practical applications. So, rapid detection method for SCC with high precision and low cost and simple operation is urgently needed for the dairy farming industry.

Near-infrared spectroscopy has been widely used in agriculture, food, medicine, and chemical industry due to its low cost, rapid, accurate, and nondestructive characteristics (Gozukara *et al*., 2022; Yang *et al*., 2022; Ciza *et al*., 2022; Douglas *et al*., 2018). Milk is suspension containing many substances of different molecular sizes. The propagation of light in milk is complex, including absorption, scattering and reflection. Researchers also tried to apply near-infrared spectroscopy to milk detection because of its analyzing advantages. Quantitative analysis models of milk components including fat and dry matter are established by near-infrared spectroscopy analysis technology (Li *et al*., 2011; Dos Santos Pereira *et al*., 2021). Near-infrared spectroscopy is also used to detect additives in milk (El-Loly *et al*., 2013). Tsenkova *et al*. (2001) established composition detection models by near-infrared spectroscopy for high and low SCC, and found that SCC affects the predictive performance of models and concluded there is an indirect relationship between SCC and near-infrared spectroscopy data from changes of compositions. To provide a rapid detection method of SCC for dairy farms, milk samples with different SCCs were collected and the feasibility of applying near-infrared spectroscopy to SCC detection in milk was researched in this paper.

Materials and Methods

Samples and data

In this study, milk samples with different SCCs were collected from the Dairy Herd Improvement Centre of Heilongjiang Animal Husbandry Service in China. There was a total of 398 effective milk samples from different dairy farms. Collected samples were divided into two parts, one part of the samples was used for measuring SCC values at the Dairy Herd Improvement Centre of Heilongjiang Animal Husbandry Service in China, while the other part was placed in a 4 ℃ insulation box and promptly sent to near-infrared spectroscopy collection laboratory. Table 1 showed sample distribution according to SCC values. SCC values of milk samples in this paper were from 6 000 to 9 085 000. The value of SCC was usually hundred thousand levels. To be convenient, millions per millilitre would be as the unit of SCC in this paper.

Table 1 Milk samples with different somatic cell count

Sample	<20	20–50	50–80	>80
Severity of mastitis	Health	Suspected	Slight	Heavier
Number of sample	150	93	98	57

Antaris II near-infrared spectrometer produced by Thermo Fisher Scientific was used to scan milk samples. The wave number range of the near-infrared spectrum is from 4 000 to 12 000 cm^{-1}, including 2 075

wavelength points. The spectral scanning resolution was 4 cm^{-1}, and an integrating sphere was used for diffuse reflection collection. Air was selected as the comparison object. Before scanning the samples, background scanning was set to 32 times, and during the experiment, the scanning frequency was set to 32 times. Each sample was scanned three times, and the average value was taken as the original near-infrared spectrum of the sample. Fig. 1 showed the average spectral data of samples with different SCCs, where the *x*-axis represented the wavelength of 4 000–12 000 cm^{-1} and the *y*-axis represented the absorbance of the samples. Observation of the graph revealed that the spectral data of milk samples with different SCCs exhibited the same trend and absorption peak, but with different absorption rates. The front and rear of the spectrum contained more noise. Therefore, data processing and modelling methods would be studied to detect SCC in milk.

Preprocessing methods

Spectral data preprocessing and outlier detection were synchronously studied in this paper. Savitzky-Golay smoothing (SG), first derivative (1D), standard normal variable (SNV), multiple scattering corrections (MSC), and combined methods were used to remove noise information from near-infrared spectral data of milk samples. Monte Carlo sampling (MCS) was used to identify abnormal samples (Cao *et al*., 2010). Kennard-Stone (KS) algorithm based on spectral variables was used to divide the sample set (Wu *et al*., 1996).

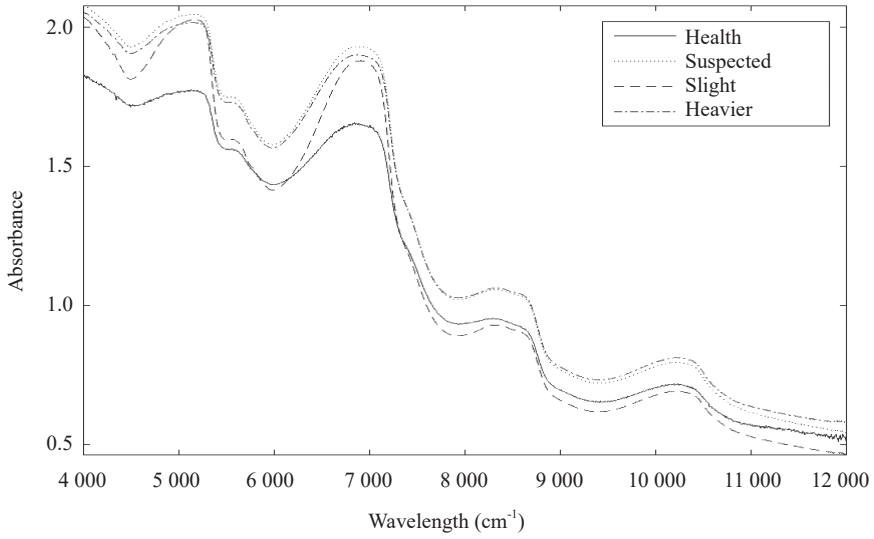

Fig. 1 Average spectra data of milk samples with different somatic cell counts

Proposed variable selection method

There were 2 075 wavelength variables in the spectral data of milk samples including lots of redundant and irrelevant informations. The variable selection method could effectively select and extract characteristic variables. Variable selection algorithm based on hybrid strategy was studied for SCC detection in milk according to the specificity of the milk samples. The hybrid strategy combined two or three variables selection methods to fully utilize the advantages of all methods and select effective variables (Yu *et al*., 2020).

Yun *et al*. (2015) summarized the variable selection methods in near-infrared spectroscopy into two categories: wavelength interval selection (WIS) and wavelength point selection (WPS). WIS methods selected continuous variable intervals, so they often

had good model interpretability. WPS methods extracted relational wavelength points as variables, and the prediction accuracy of this kind of algorithm was generally high. Based on the WIS and WPS ideas, a hybrid variable selection method the fitful SCC detection was proposed. Wavelength interval should be selected in the first step by WIS method and wavelength points would be extracted in the second step by WPS.

Synergy interval partial least squares (SiPLS) was a band selection method based on the partial least squares regression (PLSR) modelling method. This method divided spectral data into equal intervals and selected several intervals for combination. It fully considered the discontinuity of related variables in measured substance and took into account the combination effects (Sun *et al*., 2022). So SiPLS was chosen as the WIS algorithm in the first step to eliminate large number of uninformative variables, and retained strong information variables and spectral continuity.

In the second step, variables combination population analysis (VCPA) and iteration retain information variable (IRIV) were combined to extract wavelength points. The VCPA used the exponential decay function (EDF) method to quickly eliminate variables, and selected variables were usually relatively less, which might lead to the deletion of a large number of informative variables (Jiang *et al*., 2020). IRIV method fully considered the importance of each variable, which meant that it would take too much time in the face of a large number of variables (Wang *et al*., 2022). Combined VCPA and IRIV could make full use of advantages of them and make up for their shortcomings. So, VCPA made up for the problems of fewer selected variables, and eliminated variables with small contributions through EDF, resulting in easier for IRIV to screen out the optimal variable subset. So, VCPA–IRIV was chosen as the WPS method to extract variables in the second step. Competitive adaptive reweighted sampling (CARS), Bootstrapping soft, shrinkage (BOSS), genetic algorithm (GA), VCPA and IRIV could select wavelength points related to tested objects through different strategies effectively (Wang *et al*., 2017; Sun *et al*., 2020; Niazi and Leardi, 2012; Yun *et al*., 2015). To verify the effectiveness of the proposed method, these five WPS variable selection methods were compared to the hybrid variable selection method proposed in this paper. The hybrid variable selection method proposed in this paper is shown in Fig. 2.

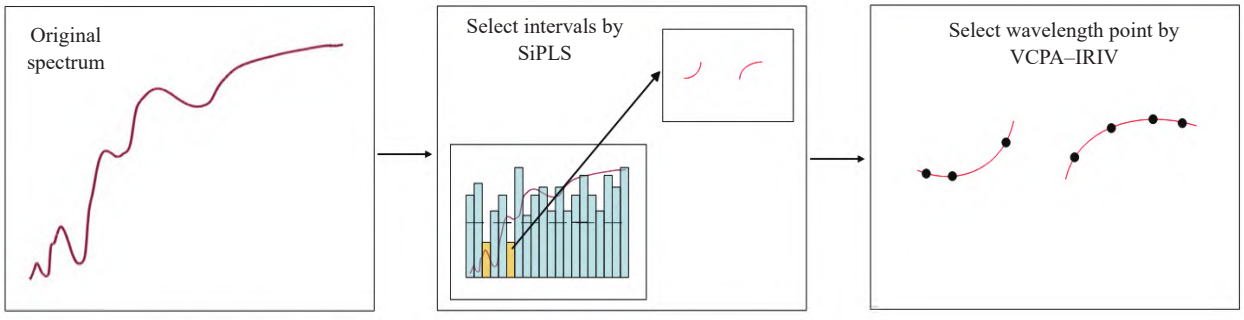

Fig. 2 Process of hybrid variable selection method for somatic cell count detection in milk

Modelling method based on ensemble learning

Ensemble learning was introduced into model construction to ensure the accuracy and stability of SCC detection model. Ensemble learning was combined with multiple models, and the error generated by a single model could be compensated by other learners. Ensemble learning methods included bagging, boosting and stacking. Bagging and boosting integrated

the same type of models, and stacking integrated heterogeneous models (Breiman, 1996). Stacking was a hierarchical model integration framework, which took the output of the base learner in the first layer as input of the meta-learner in the second layer. Different types of base learners not only had good prediction performance, but also had certain differences between each other, so that each learner could complement each other to achieve better prediction results (Tsakiridis *et al*., 2019). In this study, support vector regression (SVR), gradient boost regression tree (GBRT), PLSR, and generalized boosted regression models (GBM) were studied as base learners, and the least absolute shrinkage and selection operator (LASSO) model was meta-learner. The stacking basic learning framework model of this study is shown in Fig. 3.

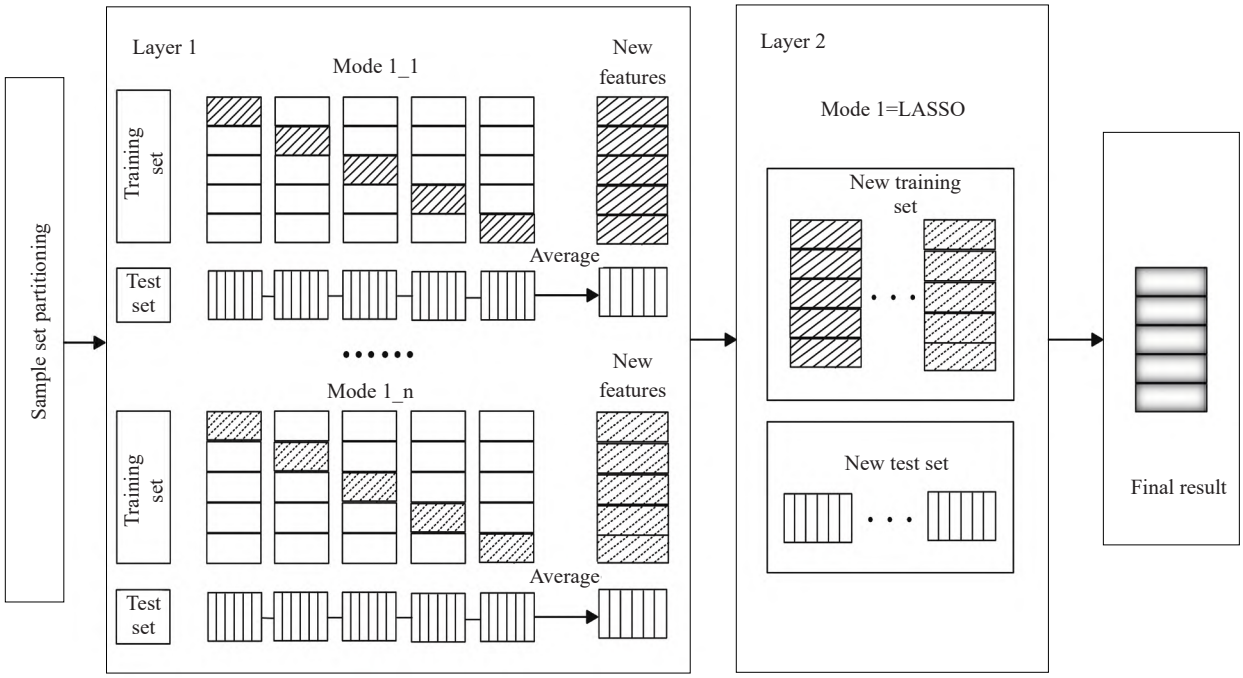

Fig. 3 Stacking ensemble learning framework for somatic cell count detection in milk

The specific step was as follows: (1) original sample data was divided into training set and test set according to the proportion, and then samples in the training set were divided into K groups with the same amount of data. (2) Base learners were used for prediction. K-1 data was used as training samples for each training, and the remaining 1 data was used for prediction. Prediction results of all base learners were used as a training set of meta-models. (3) Combined K test data to obtain new training sample data, and took the average as a new test set. (4) Put the new training set and test set into the meta-learner for training to obtain the final prediction result.

Model performance assessment

Root mean square error of cross validation (RMSECV) was used to select the optimal parameters in different models. The correlation coefficient of the calibration set (R_c^2) and root mean square error of the calibration set (RMSEC) were used to evaluate the effect of the model on the training set. The determination coefficient of the test set (R_p^2) and the root mean square error of the test set (RMSEP) were used as the evaluation indexes of the model prediction ability. A good model usually had a higher R_p^2 value and a lower RMSECP value. The formulas were as follows:

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

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

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (3)$$

In the formula, n represented the number of samples, $\hat{y}_i$ represented the true value of the ith sample, y_i represented the predicted value of the ith sample, and $\bar{y}_i$ represented the average value of the true values of all n samples.

Results

Data preprocessing

Spectral data preprocessing is shown in Fig. 4. Compared with the original spectrum, SG smoothing method removed random noise especially at the beginning and end. MSC eliminated spectral differences caused by different scattering levels due to particle inhomogeneity and different particle sizes. SNV method corrected spectral errors caused by scattering and particle size effects between samples. 1D eliminated the influence of spectral baseline drift.

Effects of PLSR models after different spectral data preprocessing methods and MCS outlier detection are shown in Table 2. SG, MSC and SNV algorithms removed noise information from the original spectrum to a certain extent. Noise information in the original spectrum was multifaceted. A combination of various spectral preprocessing methods was explored in this paper. The results showed that the combination of SG and MSC methods achieved the best result. This was because the milk sample contained a large number of macromolecular milk proteins, and macromolecular protein sizes of milk samples without homogenization treatment were different. SG smoothing could eliminate random noise of spectral data, and MSC could eliminate the influence of scattering on the spectrum caused by uneven particle size and different particle sizes. Finally, the combination of SG and MSC methods was chosen to remove noise information and seven abnormal samples were removed. Total 391 valid samples were involved in the subsequent model construction.

KS algorithm was used to divide 391 samples into training set and test set according to the ratio of 2 : 1. Results are shown in Table 3. SCC values of the training set were between 3.77 and 6.90, with a standard deviation of 0.75. SCC values of samples in the test set were between 4.11 and 6.30, and the standard deviation was 0.56. SCC range of the training set was wider than that of the test set, and sample division was reasonable.

Variable selection by hybrid strategy

Spectral intervals were selected by SiPLS in the first step. Number of intervals were 10, 20 and 30 and the number of combinations were 2, 3 and 4. PLSR models were built to choose the best parameter. Results are shown in Table 4. It could be seen that the PLSR model with the lowest RMSECV appeared at 10 interval numbers and four combination numbers. The four intervals were chosen as variables selected by SiPLS. Fig. 5 showed the four spectral intervals including 4 000–4 794 cm^{-1}, 6 398–7 197 cm^{-1}, 10 398–11 196 cm^{-1}, and 11 200–12 000 cm^{-1}. Number of wavelength points was 830, accounting for 40% of the entire spectrum. This might be due to the significant scattering effect of a large number of somatic cells in milk on light within this wavelength range, resulting in an increase in milk absorbance.

In the second step, the running times of EDF and BMS were 50 and 1 000 for the VCPA algorithm. Number of variables was set to a fixed value of 100. After optimization by VCPA, 100 selected variables were used to eliminate irrelevant and interfering variables using the IRIV method. Fig. 6 showed the results of variable selection by VCPA and IRIV. The curve in Fig. 6A was changing of variable numbers with EDF sampling running in VCPA and the point was number of selected variables after IRIV. As the number of EDF increased, the trend of RMSECV

for VCPA and IRIV is shown in Fig. 6B. Finally, 34 characteristic variables were selected. For variable intervals selected in the first step, 8, 10, 10, and 6 characteristic wavelength points were selected in the 1st, 4th, 9th, and 10th intervals by SiPLS.

In order to show the superiority of the SiPLS–VCPA–IRIV method, it was compared with other methods, and the prediction results are shown in Table 5.

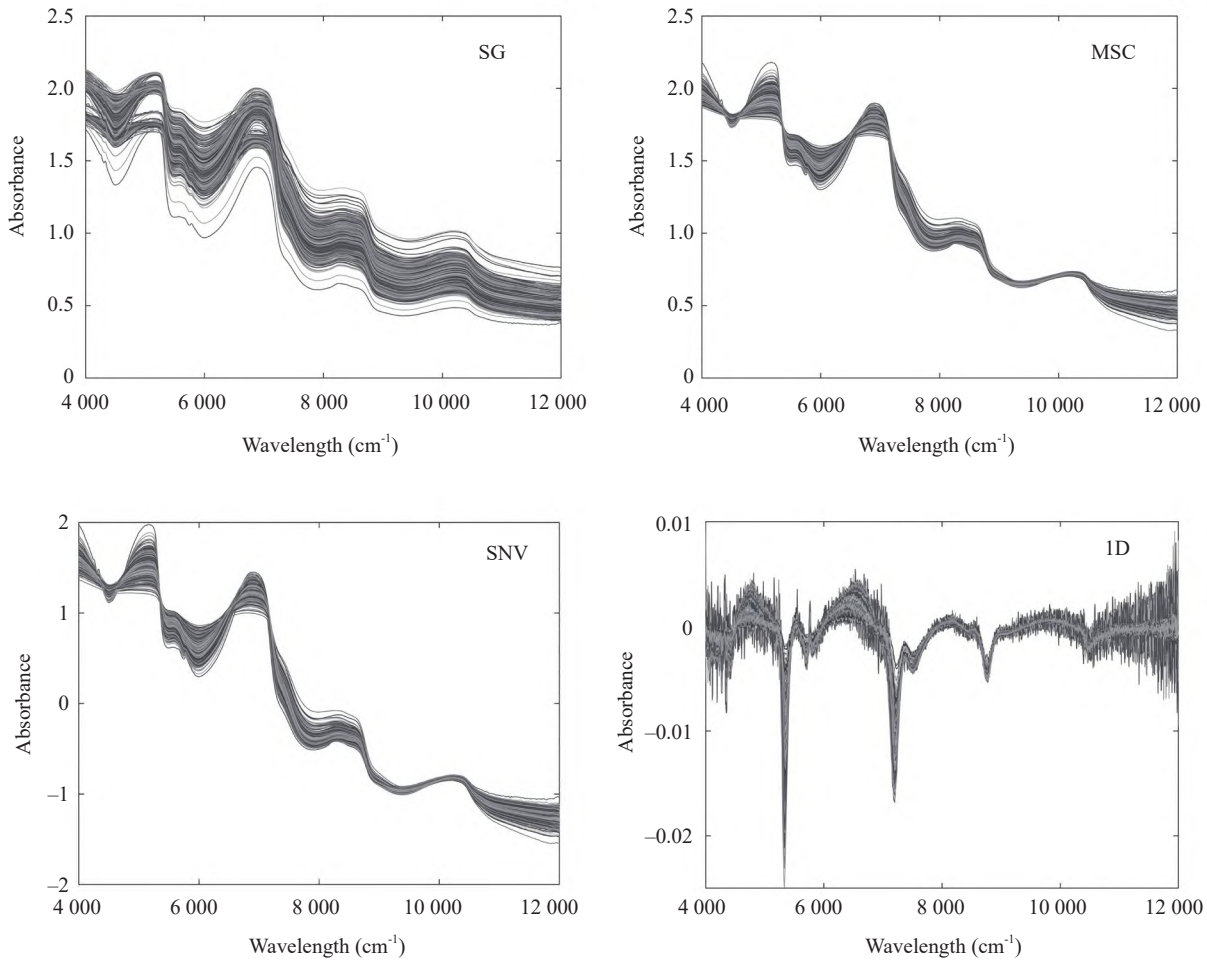

Fig. 4 Spectral data after different preprocessing methods for milk samples with different somatic cell counts

Table 2 Results of partial least squares model under different spectral preprocessing methods

Spectral preprocessing method	Number of abnormal sample	RMSEC	RMSEP
Original spectrum	9	0.50	0.45
SG	9	0.48	0.47
1D	8	0.54	0.49
MSC	8	0.47	0.50
SNV	10	0.45	0.44
MSC+SNV	8	0.46	0.46
SG+1D	9	0.49	0.47
SG+MSC	7	0.48	0.45
SG+SNV	7	0.45	0.42

Table 3 Statistical results of somatic cell count after sample set division by Kennard-Stone algorithm for milk samples

Sample set	Number of sample	Minimum of somatic cell count	Maximum of somatic cell count	Mean of somatic cell count	Standard deviation of somatic cell count
Total	391	3.77	6.90	5.38	0.70
Training set	260	3.77	6.90	5.33	0.75
Testing set	131	4.11	6.30	5.23	0.56

Table 4 Results of partial least squares models for different interval selections by synergy interval partial least squares method for somatic cell count detection in milk

Number of interval	Number of combination	Selected interval	nLV	RMSECV
10	2	1, **4**	7	0.3786
	3	1, **4, 9**	9	0.3344
	4	1, **4, 9, 10**	8	0.3059
20	2	2, **4**	7	0.4008
	3	2, **8, 15**	8	0.3465
	4	2, **8, 17, 19**	10	0.3343
30	2	3, **6**	6	0.4205
	3	2, **5, 12**	7	0.3702
	4	1, **4, 12, 22**	8	0.3396

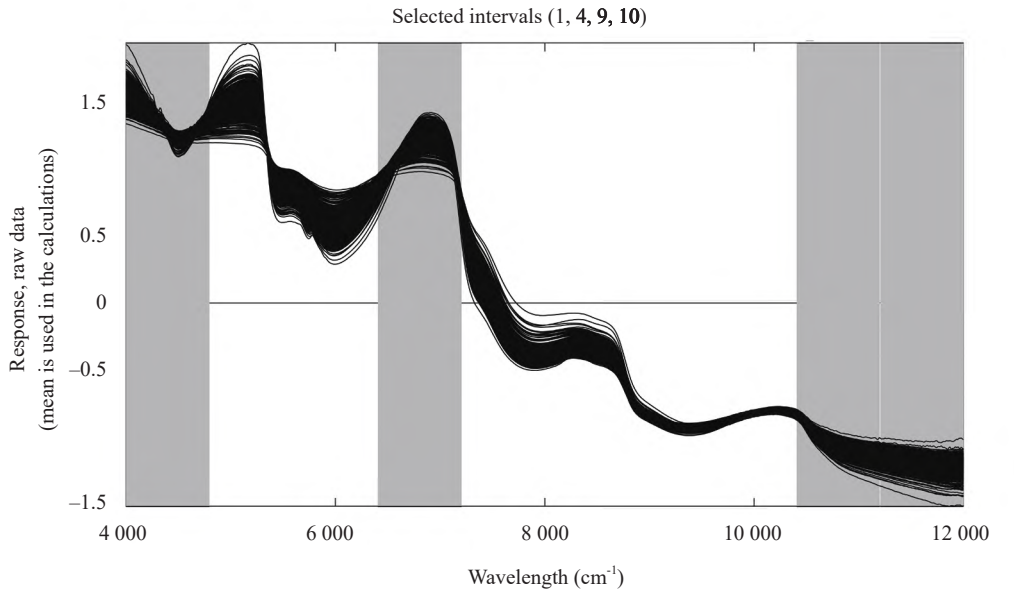

Fig. 5 Interval selection results using synergy interval partial least squares for somatic cell count detection in milk

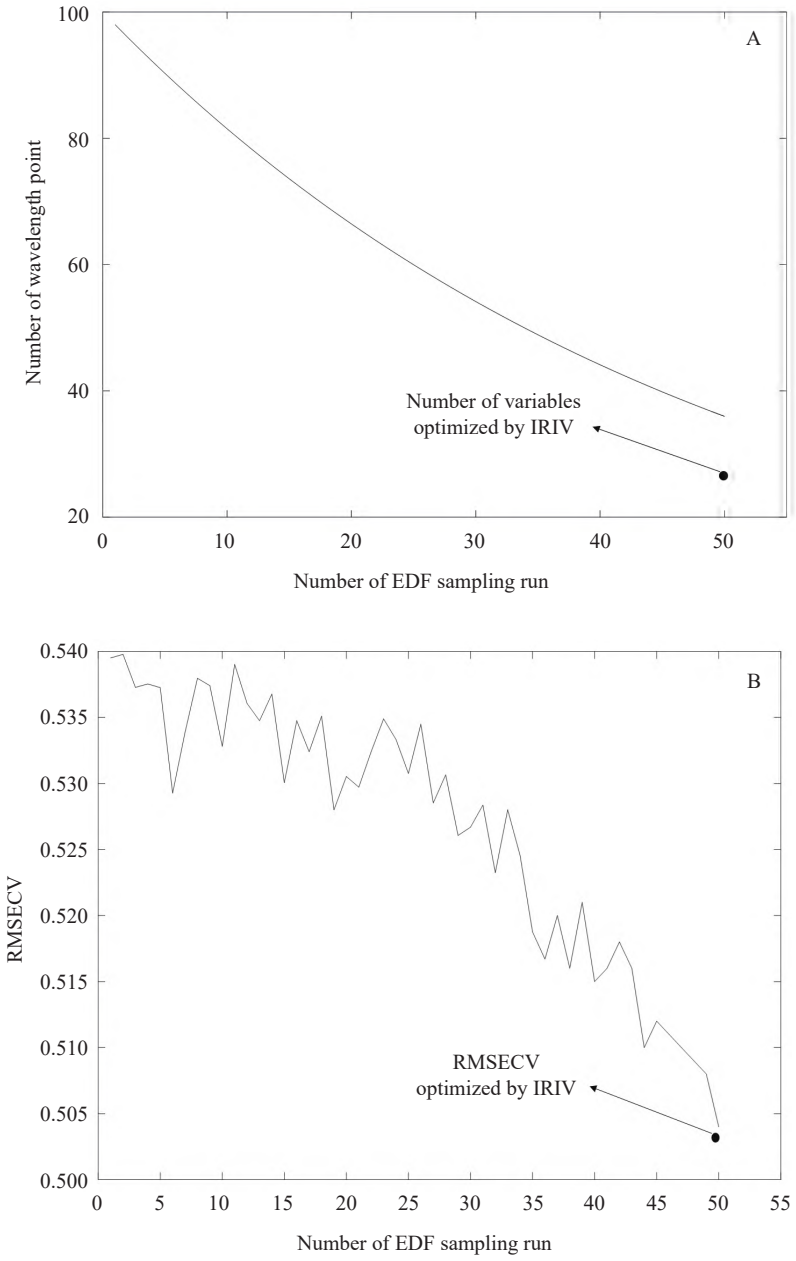

Fig. 6 Wavelength points selected results by hybrid variable selection method for somatic cell count detection in milk

The results showed that when VCPA–IRIV was not selected by SiPLS, the results were poor. The reason might be that VCPA was not effective in the selection of wavelength points when there were too many irrelevant variables, which led to the inability of the IRIV method to show its superiority. In addition, the results of the WPS method after SiPLS selection were better than those of the single WPS method.

The reason was that the frequency band selected by SiPLS retained effective information, which provided convenience for the operation of the WPS algorithm. From the above point of view, it could be seen that interval selection of SiPLS was necessary for screening useful information, and VCPA–IRIV could continue to optimize the variable space on the basis of effective information, so as to obtain better predic-

tion results.

Modeling by ensemble learning

After the wavelength selection of SiPLS–VCPA–IRIV, the wavelength was reduced from 2 075 to 34, and the number of wavelengths occupying 1.6% of the full spectrum provided better prediction results. However, on the whole, the prediction accuracy should be further improved. In view of the fact that a single model often couldn't meet the needs of performance and stability at the same time, a stacking ensemble learning algorithm was used to replace the traditional modelling method.

After training, the scatter plots of the measured and predicted values of the five base models are shown in Fig. 7.

Table 5 Results of partial least squares models with different variable selection methods for somatic cell count detection in milk

Variable selection method	Data dimension	nLV	R_c^2	RMSEC	R_P^2	RMSEP
None	2075	10	0.59	0.45	0.50	0.42
VCPA	20	9	0.61	0.46	0.55	0.40
CARS	38	10	0.58	0.46	0.50	0.45
BOSS	78	8	0.60	0.41	0.55	0.37
VCPA–IRIV	51	9	0.65	0.45	0.49	0.39
SiPLS–CARS	71	9	0.68	0.42	0.56	0.39
SiPLS–BOSS	36	9	0.66	0.40	0.53	0.38
SiPLS–GA	52	6	0.62	0.44	0.49	0.40
SiPLS–VCPA–IRIV	34	8	0.68	0.39	0.59	0.35

The predicted values of prediction models established by SVR, GBRT, PLSR, and GBM were relatively scattered. However, using the stacking ensemble model approach, the data points were concentrated at both ends of the fitted line. Therefore, the modelling method using the stacking ensemble model idea had a much higher predictive performance than the four base models of SVR, GBRT, PLSR, and GBM.

Modelling results after training are shown in Table 6. The stacking integrated model proposed in this study had a better prediction effect, prediction correlation coefficient R_p^2 was 0.66 and RMSEP was 0.30. The stacking ensemble learning model had a better prediction effect than any single model.

In order to test the robustness of the stacking ensemble algorithm, this study randomly divided the milk somatic cell count sample set 10 times according to the ratio of 2∶1, and established SVR, GBRT, PLSR, and GBM four single learner models and stacking ensemble learning models for each divided sample set.

Fig. 8 was the box plot of R_P^2 and RMSEP for different model prediction sets after 10 times of sample set division.

The medians of R_P^2 of SVR, GBRT, PLSR, GBM, and stacking integrated models were 0.62, 0.61, 0.58, 0.60, and 0.66, respectively, and the medians of RMSEP were 0.37, 0.33, 0.34, 0.34, and 0.29, respectively.

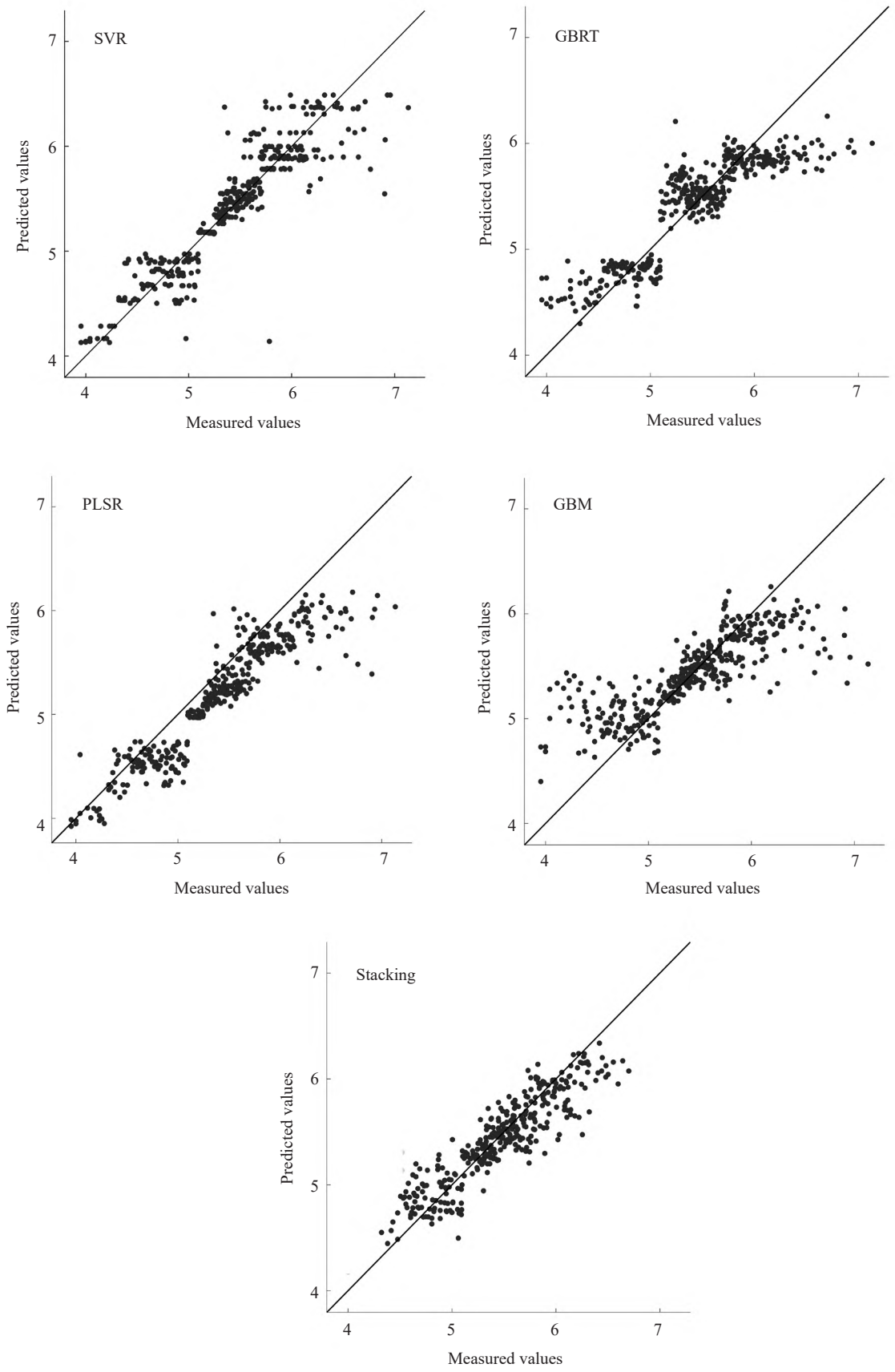

Fig. 7 Validation results of somatic cell count detection in milk based different modelling methods

Table 6 Results of single model and stacking model for somatic cell count detection in milk

Model	R_c^2	RMSEC	R_P^2	RMSEP
SVR	0.65	0.43	0.63	0.36
GBRT	0.72	0.37	0.62	0.32
PLSR	0.68	0.39	0.59	0.35
GBM	0.70	0.36	0.60	0.33
Stacking	0.72	0.34	0.66	0.30

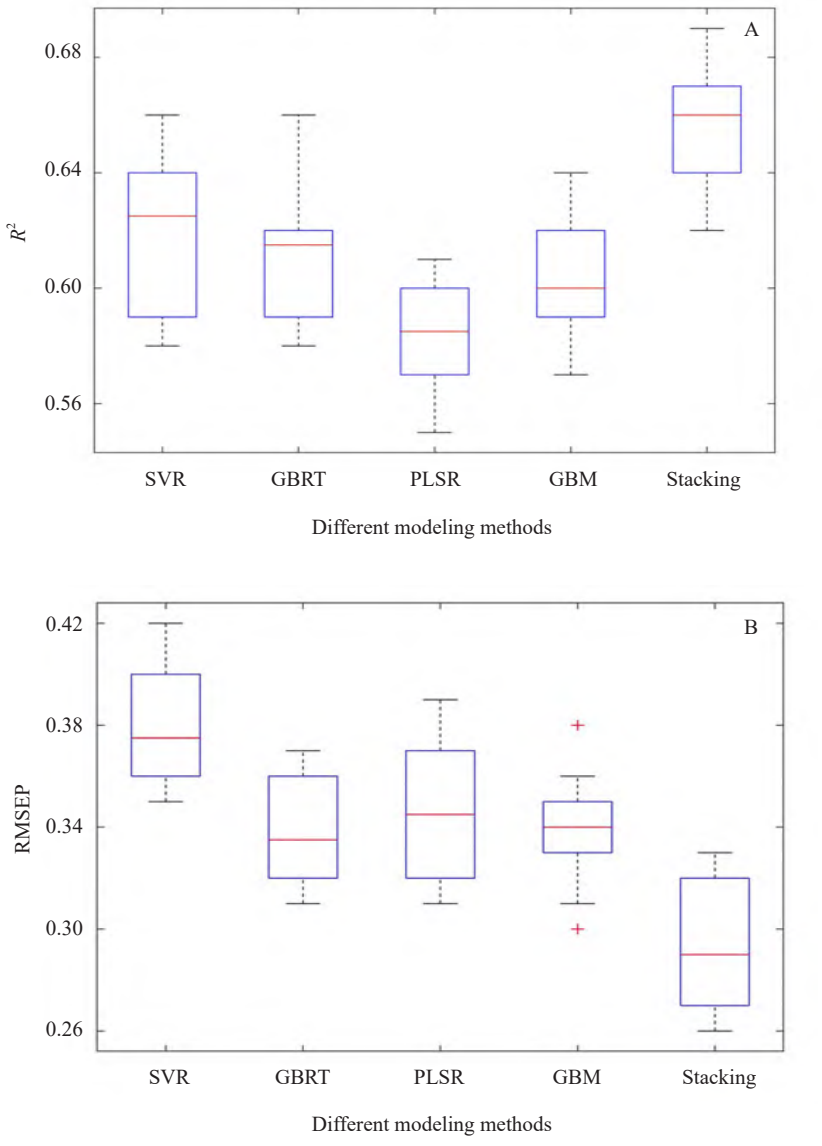

Fig. 8 Box plot of prediction results for somatic cell count detection based on single model and stacking model

Compared with the single model, the stacking integrated model had the highest R_P^2 and the lowest RMSEP. From the results, stacking integrated model could predict the somatic cell count of milk relatively stable, which proved that the method had good robustness.

Above results showed that the stacking ensemble learning model could combine the advantages of each primary learner to show better performance than any single model. It not only improved the prediction performance, but also had good robustness, so it could be stably applied to the prediction of milk somatic cell count.

Discussion

As indicator for monitoring the health of dairy cows, SCC detection was a daily task in the dairy industry. Faced frequent detection and large samples, a rapid detection method for SCC was urgently needed in practical production. In this paper, near-infrared spectroscopy analysis technology with the advantage of fast analysis was applied to SCC detection in milk. Data preprocessing, variable selection and modelling methods were researched to improve the effect of SCC detection model. But R_c^2 and R_P^2 of the final model were 0.72 and 0.66. Accuracy of SCC detection model should be further improved. An unclear essential relationship among somatic cells and spectral data and the influence of milk composition on spectral data would affect accuracy of SCC detection model. The relationship between somatic cells and spectral data of milk samples should be researched and applied it into variable selection and modelling process to improve the accuracy of detection model.

SCC detection model proposed in this paper was used to diagnose subclinical mastitis in dairy cows. The accuracy was 93%. Precision of health and subclinical mastitis were 86% and 97%. The results showed that the method proposed in this paper could meet the need of rapid diagnosis of subclinical mastitis in actual production. Near-infrared spectroscopy could be applied to SCC detection in milk to realize rapid detection of SCC for dairy farms.

Conclusions

To solve the need for SCC detection for dairy farms, the feasibility of applying near-infrared spectroscopy to SCC detection in milk was studied in this paper. Total 391 samples with different SCCs were collected and data preprocessing, variable selection and modelling methods were studied. Experimental results showed that combined SG and SNV algorithms could effectively remove noise information caused by uneven molecular in milk. A new hybrid variable selection method based on the WIS–WPS idea was proposed to extract variables from 2 075 wavelength points. SiPLS–VCPA–IRIV variable selection method can effectively extract variables. Number of variables was reduced 98.4%. To improve the generalization ability of the model, an ensemble learning strategy was applied to model building. Based on the stacking idea, SVR, GBRT, PLSR, and GBM were used as base learners, and LASSO was used as a meta-learner. Compared with the single method, SCC detection model based on an ensemble learning strategy had a better effect. Its R_c^2 was 0.72 and R_P^2 was 0.66. The accuracy of diagnosis of subclinical mastitis by the model was 93%. Near-infrared spectroscopy could be applied to SCC detection in milk to realize rapid detection of SCC for dairy farms. It was helpful to realize the rapid, non-destructive and stable detection of SCC in milk, and provided an effective method for animal husbandry production process management and quality control.

References

Albenzio M, Caroprese M. 2011. Differential leucocyte count for ewe milk with low and high somatic cell count. *Journal of Dairy*

Research, **78**(1): 43–48.

Ashraf A, Imran M. 2018. Diagnosis of bovine mastitis: from laboratory to farm. *Tropical Animal Health and Production*, **50**(6): 1193–1202.

Breiman L.1996. Stacked regressions. *Machine Learning*, **24**(1): 49–64.

Cao D S, Liang Y Z, Xu Q S, *et al.* 2010. A new strategy of outlier detection for QSAR/QSPR. *Journal of Computational Chemistry*, **31**(3): 592–602.

Ciza P H, Sacre P Y, Waffo C, *et al.* 2022. Comparison of several strategies for the deployment of a multivariate regression model on several handheld NIR instruments. Application to the quality control of medicines. *Journal of Pharmaceutical and Biomedical Analysis*, **215**: 114755. doi.org/10.1016/j.jpba.2022.114755.

Dos Reis C B M, Barreiro J R, Moreno J F G, *et al.* 2011. Evaluation of somatic cell count thresholds to detect subclinical mastitis in Gyr cows. *Journal of Dairy Science*, **94**(9): 4406–4412.

Dos Santos Pereira E V, de Sousa Fernandes D D, de Araújo M C U, *et al.* 2021. *In-situ* authentication of goat milk in terms of its adulteration with cow milk using a low-cost portable NIR spectrophotometer. *Microchemical Journal*, **163**: 105885. doi.org/10.1016/j.microc.2020.105885.

Douglas R K, Nawar S, Alamar M C, *et al.* 2018. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. *Science of the Total Environment*, **616**: 147–155.

El-Loly M M, Mansour A I, Ahmed R O. 2013. Evaluation of raw milk for common commercial additives and heat treatments. *Internet Journal of Food Safety*, **15**(10): 7–10.

Gozukara G, Akça E, Dengiz O, *et al.* 2022. Soil particle size prediction using vis-NIR and pXRF spectra in a semiarid agricultural ecosystem in Central Anatolia of Türkiye. *Catena*, **217**: 106514. doi: 10.1016/j.catena.2022.106514.

Halasa T, Huijps K, Østerås O, *et al*. 2007. Economic effects of bovine mastitis and mastitis management: a review. *Veterinary Quarterly*, **29**(1): 18–31.

He W, Ma S, Lei L, *et al.* 2020. Prevalence, etiology, and economic impact of clinical mastitis on large dairy farms in China. *Veterinary Microbiology*, **242**: 108570. doi: 10.1016/j.vetmic.2019.108570.

Jiang H, Xu W, Ding Y, *et al.* 2020. Quantitative analysis of yeast fermentation process using Raman spectroscopy: comparison of CARS and VCPA for variable selection. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, **228**: 117781. doi: 10.1016/j.saa.2019.117781.

Keisler D H, Andrews M L, Moffatt R J, *et al.* 1992. Subclinical mastitis in ewes and its effect on lamb performance. *Journal of Animal Science*, **70**(6): 1677–1681.

Li X, Wang J, Huang Y, *et al.* 2011. Determination of fat, protein and DM in raw milk by portable short-wave near infrared spectrometer. *Spectroscopy and Spectral Analysis*, **31**(3): 665–668.

Ma Y, Ryan C, Barbano D M, *et al.* 2000. Effects of somatic cell count on quality and shelf-life of pasteurized fluid milk. *Journal of Dairy Science*, **83**(2): 264–274.

Niazi A, Leardi R. 2012. Genetic algorithms in chemometrics. *Journal of Chemometrics*, **26**(6): 345–351.

Sun D M, Huan K W. 2020. Research on near infrared spect-roscopy analytical methods of moisture content in wheat based on BOSS. *Journal of Changchun University of Science and Technology* (*Natural Science Edition*), **43**(5): 1–6.

Sun H, Kong F, Xiu C, *et al.* 2022. A progressive combined variable selection method for near-infrared spectral analysis based on three-step hybrid strategy. *Journal of Spectroscopy*, **2022**: 2190893. doi: 10.1155/2022/2190893.

Tsakiridis N L, Tziolas N V, Theocharis J B, *et al.* 2019. A genetic algorithm-based stacking algorithm for predicting soil organic matter from vis–NIR spectral data. *European Journal of Soil Science*, **70**(3): 578–590.

Tsenkova R, Atanassova S, Ozaki Y, *et al.* 2001. Near-infrared spectroscopy for biomonitoring: influence of somatic cell count on cow's milk composition analysis. *International Dairy Journal*, **11**(10): 779–783.

Urbanová E, Sedinova V, Skarda J, *et al.* 1985. Use of the Synpor membrane filter for the separation and determination of the number of somatic cells in the milk of dairy cows using the indole DNA filtration method. *Veterinarni Medicina*, **30**(7): 409–418.

Wang S, Sun J, Fu L, *et al.* 2022. Identification of red jujube varieties based on hyperspectral imaging technology combined with CARS-IRIV and SSA-SVM. *Journal of Food Process Engineering*, **45**(10): e14137. doi: 10.1111/jfpe.14137.

Wang Y, Jiang F, Gupta B B, *et al.* 2017. Variable selection and

optimization in rapid detection of soybean straw biomass based on CARS. *IEEE Access*, **6**: 5290–5299.

Wu W, Walczak B, Massart D L, *et al.* 1996. Artificial neural networks in classification of NIR spectral data: design of the training set. *Chemometrics and Intelligent Laboratory Systems*, **33**(1): 35–46.

Yang Q, Xie C, Luo K, *et al.* 2022. Rational construction of a new water soluble turn-on colorimetric and NIR fluorescent sensor for high selective Sec detection in Se-enriched foods and biosystems. *Food Chemistry*, **394**: 133474. doi: 10.1016/j.foodchem.2022.133474.

Yu H D, Yun Y H, Zhang W, *et al.* 2020. Three-step hybrid strategy towards efficiently selecting variables in multivariate calibration of near-infrared spectra. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, 224. doi: 10.1016/j.saa.2019.117376.

Yun Y H, Wang W T, Deng B C, *et al.* 2015. Using variable combination population analysis for variable selection in multivariate calibration. *Analytica Chimica Acta*, **862**: 14–23.

Zhang X, Xue H, Xiaojing G. 2018. Milk somatic cells recognition based on Gray-Scale difference statistics. *MATEC Web of Conferences*, **173**: 03065. doi: 10.1051/matecconf/201817303065.