Effect of variable selection and rapid determination of total tea polyphenols contents in Fuzhuan tea by near-infrared spectroscopy

Jing-Xue Liu\textsuperscript{ab}, Jia-Ying Xin\textsuperscript{ac}, Ting-Ting Gao\textsuperscript{b}, Feng-Lin Li\textsuperscript{b} and Xie Tian\textsuperscript{b}

\textsuperscript{a}Key Laboratory for Food Science & Engineering, Harbin University of Commerce, Harbin, Heilongjiang, China; \textsuperscript{b}College of Food Engineering, Jilin Agricultural Science and Technology University, Jilin, Jilin, China; \textsuperscript{c}State Key Laboratory for Oxidation & Selective Oxidation, Lanzhou Institute of Chemical Physics, Chinese Academy of Sciences, Lanzhou, Gansu, China

\section*{ABSTRACT}
This study attempted to measure the total polyphenols contents in Fuzhuan tea by near-infrared (NIR) spectroscopy coupled with an appropriate multivariate calibration method. Partial least squares (PLS), synergy interval PLS (si-PLS), and genetic algorithm-based PLS (ga-PLS) were carried out comparatively to calibrate regression models. The root mean square error of prediction (RMSEP), determination coefficient (R\textsuperscript{2}), and \textit{p}-value between the true and predicted values of prediction set were used to evaluate the performance of the final model. The ga-PLS model showed the best performance compared with the PLS and si-PLS models. The optimal model obtained Rp\textsuperscript{2} = 0.9996 and RMSEP = 0.0488 for the prediction set using only 37 spectral data points. No significant difference was observed between the true and predicted tea polyphenol contents in the prediction set (\textit{P} > 0.05). NIR spectroscopy together with the ga-PLS algorithm can be used to rapidly predict the total polyphenol contents in Fuzhuan tea.

\section*{1. Introduction}

Fuzhuan tea, a unique type of dark tea, is mainly produced in Anhua county, Hunan province, China, for more than 200 years history (Zhou et al., 2021). Fermentation is among the most important processes in Fuzhuan tea production. During fermentation, the raw tea materials undergo a series of biochemical reactions, in which tea polyphenols are the main reactive components, forming unique quality characteristics of Fuzhuan tea soup, which has a bright color and thick mellow taste (Ghosh et al., 2021). Fuzhuan tea contains many beneficial components, such as tea polyphenols, theabrownin, and thearubigin. When tea soup is absorbed by the human body, it can decompose fat (Liu et al., 2019), relax the stomach (Zhu et al., 2015), resist oxidation (Zhang & Fang, 2018), and decrease blood sugar (Chen et al., 2021), among other functions. Hence, Fuzhuan tea is popular with consumers and has become indispensable in their daily life.

Tea polyphenols are greatly interesting owing to their beneficial medicinal properties (Prawira-Atmajja et al., 2022; Samynathan et al., 2021). A large number of studies have confirmed that polyphenols in tea could improve health. Currently, a study has shown that antioxidants in polyphenols from tea might play an important role in the prevention of cardiovascular disease (Maleki et al., 2019), chronic gastritis (Jantan et al., 2021), and some cancers (Ahmed et al., 2019; Ceci et al., 2018; Coscaras et al., 2019). Furthermore, the characteristics of astringency and bitterness of brewed tea are caused by polyphenol compounds (Rocchetti et al., 2018). Therefore, the tea polyphenol content is an index for evaluating the quality of Fuzhuan tea. Generally, the Folin phenol chemical method is employed to measure tea polyphenol contents (Punyasiri et al., 2015). Nevertheless, this method is not only time-consuming, has high requirements for analysts, and laborious, but also pollutes the...
peripheral environment. Therefore, developing a fast and non-destructive method to determine tea polyphenol contents is necessary.

In recent years, near-infrared (NIR) spectroscopy, as a fast, accurate, and non-destructive method, can be employed to substitute the traditional chemical analysis methods. NIR spectroscopy, as a powerful analytical tool, was widely applied in agricultural (Xue & Jian, 2014; Zhou et al., 2015), petrochemical (Wu et al., 2014), textile (Tavanaie et al., 2015), and pharmaceutical industries (Blanco & Peguero, 2010; Lee et al., 2011). In addition, NIR spectroscopy has been extensively used in the tea industry in recent years, for example, to analyze caffeine and free amino acids, determine the origins of tea varieties, and evaluate the quality of tea (Li et al., 2021; Rehman et al., 2020). The tea nutrition model was calibrated by backward interval partial least squares (bi-PLS), back propagation-artificial neural networks (BP-ANN) algorithms, and other pattern recognition systems (Chen et al., 2015; Zhu et al., 2020). However, in terms of determining tea polyphenol contents in Fuzhuan tea, the use of NIR spectroscopy technology combined with multivariate data analysis methods has yet to be reported.

In these studies before calibrating the NIR spectroscopy models, many spectral preprocessing methods have been established to decrease the effects of variations in the spectral data, which are unrelated to chemical variations in the samples. These methods could usually significantly enhance the prediction effect of the calibration model. However, some selected spectral regions might not contain any useful information about chemical changes in the sample (Sheehan et al., 2019). Therefore, spectral data reflecting the chemical information of samples are not selected accurately. In fact, selecting appropriate spectral regions to obtain the best performance is a major problem in multivariate data analysis. Recently, a large number of theoretical and experimental evidence indicated that spectral region selection could remarkably enhance the performance of these calibration models and produce the lowest prediction error (Liu et al., 2020; Silva et al., 2016).

Therefore, developing a method to select the best variable set is important. The partial least squares (PLS) method, as a new multivariate statistical data analysis method, combines the merits of principal component analysis, canonical correlation analysis, and multiple linear regression analysis (Mohammed et al., 2020). The main research objective of this method is to determine the relevance between multivariate regression modeling and multiple independent variables. The PLS regression method is more effective and robust, especially when the internal variables indicate highly linear correlation.

The synergy interval PLS (si-PLS) method was proposed by Lin et al. (2019) for the selection of several intervals of spectral data, in which the data set is split into various intervals (10–25 intervals), and all possible PLS model combinations with two, three, or four intervals are calculated (Lin et al., 2019). The si-PLS prediction model is best when the root mean square error of cross-validation (RMSECV) value is lowest, and the best combination of spectral intervals is selected. The si-PLS method can reduce spectral noise effectively and enhance the prediction accuracy of models.

Genetic algorithm PLS (ga-PLS) is an optimization method for simulating the evolutionary mechanism of biological species competition selection based on biological evolution theory (Yang et al., 2017). This method achieves iterative optimization of a group by applying genetic manipulation to individuals in the group in accordance with a fitness function. The ga-PLS algorithm is applied to precisely select spectral data points for optimization models. This calibrating model has the best prediction ability when the RMSECV value is lowest.

In this research, after obtaining NIR spectra and selecting the spectral variables, prediction models for tea polyphenol contents were established using the PLS, si-PLS, and ga-PLS methods. This study systematically explored the different steps to be applied in each calibration model. The numbers of PLS factors and region intervals or spectral data points were optimized according to the RMSECV of the calibration set. The root mean square error of prediction (RMSEP), the determination coefficient (R²), and the P-value of the prediction set were used to evaluate the performance of the final model.

2. Materials and methods

2.1. Sample preparation

A total of 128 Fuzhuan tea samples, processed between May 2019 and September 2019, were obtained from Anhua tea market. All samples were crushed and passed through a 60-mesh sieve. The samples were randomly divided into two sets in a 3:1 ratio, resulting in 96 calibration and 32 prediction set samples, which were used to establish the calibration models and prediction models, respectively.

2.2. Determination of tea polyphenols

The tea polyphenol contents were determined using the Folin-Ciocalteu method (Zhou et al., 2022). After grinding the tea samples, tea polyphenols were distilled using 70% methanol and a water bath at 70°C. The – OH groups of tea polyphenols were oxidized by the Folin-Ciocalteu reagent (0.2 mol/L) and reacted for 5 min in the dark at 25°C. Subsequently, 2 mL of saturated sodium carbonate solution (75 g/L) was added, and then incubated in the dark for 2 h at 25°C. Finally, the absorbance was then measured using a spectrophotometer at a wavelength of 765 nm. The standard curve was drawn using gallic acid as the standard correction substance to determine the tea polyphenol contents.

2.3. Spectra collection

The NIR spectra of the sample was collected by a Thermo Antaris II Fourier transform (FT) NIR spectrometer, which was coupled with an InGaAs detector, a quartz halogen lamp, and an integrating sphere accessory, and the acquisition method was diffuse reflection method. The samples were placed in the special sample cup (ø 30 mm) designed for this experiment. For each sample, Fuzhuan tea (10 g) was placed into the sample cup following the procedure specified by the manufacturer. During acquisition, the spectrometer parameters are set as follows: spectral range of 10,000–4,000 cm⁻¹, scanning times of 16 times, resolution of 8 cm⁻¹, and data
sampling interval of 3.865 cm$^{-1}$. Each spectrum contained 1557 wavenumber points. The background of the image should be removed from the hyperspectral image to obtain the typical spectrum of samples. The background removal operation is a mask constructed by subtracting a low reflection band from the high reflection band, and then segmenting the image through mask and wide value calculation. After image segmentation, the pixels of each sample can be easily identified as the region of interest (ROI), and then the average value with three experiments of the spectrum within the ROI is used as the spectrum of each sample for subsequent analysis. During the experiment, the indoor temperature was controlled at 25 °C.

2.4. Ga-PLS model

The genetic algorithm (ga) includes five components (encoding, population initialization, individual selection, crossover, and mutation, Figure 1). Input spectral variables were encoded into binary data: zeros and ones as chromosomes. PLS is a non-parametric regression method, which is based on the idea of high-dimensional projection. It integrates the basic functions of multiple linear regression, canonical correlation, and principal component analyses.

A simple PLS model consists of two external relationships and one internal relationship (Figure 1). Let $X \in [n \times m]$ represent an explanatory matrix, the first outer relation is derived using applying PCA to $X$, resulting in the score matrix $T \in [n \times a]$ and the loading matrix $P' \in [a \times m]$ plus an error matrix $E \in [n \times m]$, i.e. $X=T P' + E$, where “a” is the number of potential variables used to explain independent variables. Similarly, the second outer relation for $Y \in [n \times p]$ accounting for a response variable matrix can be derived by decomposing $Y$ into the score matrix $U \in [n \times a]$, the loading matrix $Q' \in [a \times p]$, and the error term $F \in [n \times p]$, i.e. $Y=UQ' + F$. The inner relation $U=BT$ is a multiple linear regression between the score matrices $U$ and $T$, in which $B$ is an $n \times n$ regression coefficient matrix measured by least square minimization. The goal of the PLS model is to minimize the norm of $F$ while maximizing the co-variance between $X$ and $Y$ by the inner relation.

2.5. Spectral preprocessing and model performances

Each spectrum was transformed into 1,557 pairs of data points. To reduce spectral noise and select the best pretreatment method, all spectra were pretreated with derivative, smoothing, and multiple scattering corrections using TQ Analyst 9.4.45, OPUS 7.0, and Matlab 7.0 software to select the best pretreatment method. The comparative results of the models showed that smoothing was the best preprocessing method. The performances of the NIR models were evaluated using the determination coefficient of cross-validation (Rc2), Rp2, RMSECV, RMSEP, factors, and $P$-value. $P$-value was computed using the paired $t$-statistic method, and it was used to test for significant differences between the true and predicted tea polyphenol contents.

RMSECV is calculated as follows:

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}$$ (1)

The equation 2 was used to calculate the RMSEP.

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

The equation 3 was used to calculate the $R^2$.

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

where $\bar{y}$ is the mean true value, $y_i$ is the true value for sample $i$, and $\hat{y}_i$ is the predicted value for sample $i$.

---

Figure 1. Ga-PLS algorithm implementation flowchart.

Figura 1. Diagrama de flujo de la implementación del algoritmo ga-PLS.

Note: $X$ is the independent variable data matrix; $Y$ is the dependent variable data matrix; $T$ and $U$ are the principal component matrix extracted from $X$ and $Y$; $P'$ and $Q'$ are the load matrix; $E$ and $F$ are the residual matrix; $n$, $m$, $a$, and $p$ are the number of rows and columns in the matrices.

Nota: $X$ es la matriz de datos de la variable independiente; $Y$ es la matriz de datos de la variable dependiente; $T$ y $U$ son la matriz de componentes principales extraída de $X$ y $Y$; $P'$ y $Q'$ son la matriz de carga; $E$ y $F$ son la matriz residual; $n$, $m$, $a$ y $p$ son el número de filas y columnas de las matrices.
3. Results and discussion

3.1. Tea polyphenols contents

The tea polyphenol contents were determined using the standard method, with the results shown in Table 1. Table 1 shows that the range, mean, and standard deviation (SD) values of the calibration set were 3.40%–11.11%, 6.63%, and 2.42, respectively, whereas the corresponding values of the prediction set were 3.54%–10.42%, 5.98%, and 2.05, respectively. Therefore, the range of tea polyphenol contents in the prediction set was included in the standards of the calibration set, thereby providing a foundation for more accurate prediction of the tea polyphenol contents. The sample classification method was also proven to be correct.

3.2. NIR spectra

The mean NIR spectra of 128 Fuzhuan tea samples were the same as the major information at the spectral region of 4,500 – 5,400 cm⁻¹, 6,500 – 7,200 cm⁻¹, and 8,200 – 8,800 cm⁻¹ (Figure 2). Spectral features were interpreted as an association in the 4,500–5,400 cm⁻¹ region, in which the water molecules are present (4,500–5,400 cm⁻¹) and O – H combinations (6,500–7,200 cm⁻¹) are present. In addition, the small peak observed between 8,200 and 8,800 cm⁻¹ can be correlated with the presence of sugars, which are associated with the second and third overtones of the C – H bond (Wang et al., 2018). Similar results were reported in research performed with intact doyvalis fruit, with the NIR spectra being dominated by the presence of water with overtone bands of OH bonds and combinations (Mateus et al., 2018). Therefore, NIR spectra were preprocessed to develop the PLS models (Figure 2).

3.3. Results of NIR spectroscopy models

3.3.1. Results of PLS model

PLS is an effective linear modeling method, and the number of factors is important. The NIR spectroscopy – PLS prediction model of tea polyphenol content was established using the full spectra, with R² and RMSECV values of 0.9177 and 0.0849%, respectively, when 10 factors were included in the calibration model. Therefore, the optimal number of PLS factors was 10. The scatter plot in Figure 3 shows a correlation between the true measured and predicted NIR values in the prediction set obtained using the PLS model. As shown in Figure 3, when the PLS model performance was evaluated using samples in the prediction set, the R² and RMSECV values were 0.9037 and 0.0887%, respectively. Furthermore, the P-value between the true and predicted values in the prediction set samples was 0.24 (P > 0.05), showing no significant difference. The selected spectral regions were [2 3 7]. The PLS method can be used to predict the tea polyphenol accurately. However, the prediction effect of the PLS model can be further improved.

3.3.2. Results of si-PLS models

The si-PLS algorithm can divide the whole NIR spectral data set into multiple intervals (10 to 25) and calculate all possible PLS model combinations of two, three, or four intervals. According to the RMSECV value in the calibration set, the selection of the spectral region most related to tea polyphenols was optimized. The si-PLS model that provided the lowest RMSECV contained NIR information specific to the tea polyphenols based on optimization results. This model allowed the calculation load to be reduced (Li et al., 2020; Lin et al., 2021).

Table 2 shows the results of si-PLS calibration models when the whole spectra were split into different numbers of intervals. When the full spectra were divided into 16 subregions and the factor number was 7, the results of the calibration model were the best and RMSECV value was the lowest, at 0.0617%. Meanwhile, the largest R² value in the calibration set model was 0.9597. The selected spectral regions were [2 3 7 10], and the corresponding spectral wavelength ranges were 4377.62–4751.71, 4755.63–5129.70, 6262.72–6633.94, and 7386.02–7756.31 cm⁻¹, respectively (Figure 2).

The scatter plot in Figure 4 shows a correlation between the true measured and NIR predicted values in the prediction set obtained using the si-PLS method. When the performance of the si-PLS calibration model was evaluated using the samples in the prediction set, the RMSEP and R² values were 0.0620% and 0.9583, respectively. Furthermore, the P-value between the true and predicted values in the prediction set samples was 0.65 (P > 0.05), showing no significant difference. The si-PLS method can be used to predict the tea polyphenol accurately. By comparing the results in Figures 3 and 4, the prediction results of the best si-PLS model were better than those of the best PLS model. However, the selected spectral intervals might still contain some noise information, indicating that the spectral feature data information can be further screened.

Table 1. Classification and tea polyphenols contents of all samples.

| Set          | Range (%) | Mean (%) | SD  | CV  |
|--------------|-----------|----------|-----|-----|
| Calibration  | 3.40–11.11| 6.63     | 2.42| 1.37|
| Prediction   | 3.54–10.42| 5.98     | 2.05| 1.06|

Note: SD is the standard deviation; CV is the coefficient variation.

Nota: SD es la desviación estándar; CV es el coeficiente de variación.
calibration model was evaluated using samples in the prediction set, the RMSEP and Rp2 values were 0.0611% and 0.9996, respectively (Figure 7). The P-value between the true and predicted values in the prediction set was 1.46 ($P > 0.05$), showing no significant difference. The selected spectral regions were [3 7 10]. Therefore, the gs-PLS method can be used to predict the tea polyphenols more accurately.

The 37 spectral data points selected by the genetic algorithm were 4065.21, 4335.19, 4389.19, 4551.18, 5654.26, 5820.11, 5839.40, 5908.82, 5947.39, 5997.53, 6032.24, 6055.39, 6741.92, 6954.05, 6981.05, 7409.17, 7482.45, 7679.15, 7729.29, 8072.56, 8103.42, 8176.70, 8427.40, 8616.39, 8631.82, 8670.39, 8693.53, 8724.38, 8789.95, 8932.66, 8994.37, 9052.22, 9102.36, 9144.79, 9191.07, 9191.07,

### Table 2. Results of si-PLS calibration model selected different spectral regions.

| Number of intervals | PLS factors | Selected intervals | RMSECV (%) |
|---------------------|-------------|--------------------|------------|
| 10                  | 10          | [1 3]              | 0.0644     |
| 11                  | 10          | [1 3 6 10]         | 0.0650     |
| 12                  | 9           | [5 6 7 11]         | 0.0645     |
| 13                  | 9           | [6 7 10 11]        | 0.0649     |
| 14                  | 10          | [1 6 7 10]         | 0.0647     |
| 15                  | 10          | [1 6 9]            | 0.0643     |
| 16                  | 7           | [2 3 7 10]         | 0.0641     |
| 17                  | 10          | [1 6 9 11]         | 0.0644     |
| 18                  | 9           | [3 9 12 16]        | 0.0637     |
| 19                  | 8           | [8 10 16 18]       | 0.0643     |
| 20                  | 8           | [3 9 13 14]        | 0.0625     |
| 21                  | 9           | [3 9 12 20]        | 0.0635     |
| 22                  | 8           | [9 12 18]          | 0.0638     |
| 23                  | 10          | [1 2 6]            | 0.0638     |
| 24                  | 10          | [1 4 8 13]         | 0.0639     |
| 25                  | 9           | [4 8 16 21]        | 0.0628     |

Si-PLS: Synergy interval partial least squares; RMSECV: Root mean square error of cross validation.

Si-PLS: Mínimos cuadrados parciales del intervalo de sinergia; RMSECV: Error cuadrático medio de la validación cruzada.

### 3.3.3. Results of ga-PLS model

When using the ga-PLS method, the genetic algorithm is applied to select spectral data points for the optimization models. When the RMSECV value of the tea polyphenol model is lowest, the calibration model has the best prediction ability.

As shown in Figure 5, as the number of selected data points gradually increased, the RMSECV value rapidly decreased from 1.0223% to 0.0609%. When 37 data points were selected (Figure 6) (accounting for 2.4% of the 1,557 total spectral data points), the lowest RMSECV value was 0.0609%. Furthermore, the R² value was 0.9998, and the best number of factors in the calibration set was 10. When the performance of the ga-PLS calibration model was evaluated using samples in the prediction set, the RMSEP and Rp2 values were 0.0611% and 0.9996, respectively (Figure 7). The P-value between the true and predicted values in the prediction set was 1.46 ($P > 0.05$), showing no significant difference. The selected spectral regions were [3 7 10]. Therefore, the ga-PLS method can be used to predict the tea polyphenols more accurately.
Figure 5. The relationship between RMSECV and the number of selected variables by ga-PLS algorithm.

Figura 5. Relación entre el RMSECV y el número de variables seleccionadas por el algoritmo ga-PLS.

Figure 6. The best 37 spectra data points selected by ga-PLS algorithm.

Figura 6. Los mejores 37 puntos de datos del espectro seleccionados por el algoritmo ga-PLS.

Figure 7. True values versus NIR predicted values by ga-PLS in prediction set.

Figura 7. Valores reales frente a los valores de predicción NIR por ga-PLS en el conjunto de predicción.
Table 3. Performance of PLS, si-PLS, and ga-PLS models using NIR spectra of Fuzhuan tea for Folin-Ciocalteu method determination.

| Model   | Performance | LV | Tea polyphenols contents(%) |
|---------|-------------|----|------------------------------|
| PLS     | RMSEP (0.0887) | Rp2 (0.9037) | 19 | 5.57 |
| si-PLS  | RMSECV (0.0617) | R2 (0.9597)  | 21 | 5.81 |
| ga-PLS  | RMSEP (0.0611) | R2 (0.9996)  | 13 | 6.12 |

9503.48, and 9958.60 cm\(^{-1}\). The distribution of data points shows that 33 spectral data points were distributed in the long-wave region (5,000–10,000 cm\(^{-1}\)) of the NIR spectrum, accounting for 89.19% of the 37 total spectral data points, whereas four spectral data points were distributed in the short-wave region (4,000–4,999 cm\(^{-1}\)) of the NIR spectrum, accounting for 10.81% of the 37 total spectral data points. This result was mainly because spectral information was abundant in the long-wave NIR region (cm\(^{-1}\)), representing the majority of spectral information for Fuzhuan tea samples. On the contrary, the short-wave NIR region (cm\(^{-1}\)) of Fuzhuan tea samples had less spectral information. Therefore, when the spectral data points were optimized by the genetic algorithm, they were mainly located in the long-wave NIR region (cm\(^{-1}\)).

3.3.4. Partial least squares: Tea polyphenols contents

The performance of the models for tea polyphenols contents is shown in Table 3. Considerable variations were observed among the PLS, si – PLS, and ga – PLS regression models, with the best prediction model for tea polyphenol contents obtained by ga – PLS (Table 3). To obtain these results, NIR spectra were preprocessed by the first derivative of Savitzky – Golay, which has three windows and eliminates two outliers detected by Hoteling T\(_2\) test (P < 0.05). A ga – PLS model with 13 latent variables (LV), lower errors, and higher determination coefficients can be established using this procedure. Moreover, the tea polyphenols contents predicted by models PLS, si – PLS, and ga – PLS were 5.57%, 5.81%, and 6.02%, respectively. The results imply that the ga-PLS model showed the best performance compared with the PLS and si-PLS models.

4. Discussion

PLS is a robust and widely used method (Li et al., 2020). Therefore, the PLS method is preferred to establish a fast and non-destructive prediction model for tea polyphenol contents in Fuzhuan tea. A comparison of the prediction results of PLS, si-PLS, and ga-PLS models indicated that the ga-PLS model was the best, followed by the si-PLS model. The model with the worst performance was the PLS model possibly due to the full wavelength (involving 1,557 spectral data points) used to establish the PLS prediction model. Noise information in the spectra affected the prediction efficiency, resulting in a poor prediction ability.

Based on the PLS algorithm, the si-PLS method optimized the spectral intervals closely related to tea polyphenols. When these characteristic spectral intervals (four spectral intervals, approximately 389 spectral data points) were used for modeling, some spectral noise information were eliminated and the modeling computation was reduced. In addition, the prediction ability of the model was improved. Therefore, the results of the si-PLS model were better than those of the PLS model. The ga-PLS model was superior. When the genetic algorithm was applied, the most useful spectral data points (37 spectral data points) were extracted for modeling, thereby eliminating more noise information. Therefore, the performance of the ga-PLS model was the best. In this study, about half of the best spectral intervals selected by the si-PLS method were distributed in the NIR region of 4,000–5,000 cm\(^{-1}\) (approx. 194 spectral data points), whereas only four spectral data points selected by the ga-PLS method were distributed in the NIR region of 4,000–5,000 cm\(^{-1}\). The reason for this inconsistency in the spectral data points selected by the two methods (si-PLS and ga-PLS) requires further study.

In future studies, we will attempt to use other methods to optimize the feature spectral information related to tea polyphenols, and establish prediction models using non-linear algorithms, such as BP-ANN, because the components in Fuzhuan tea are very complex and the NIR spectra of other substances might affect tea polyphenol modeling. Finally, by comparing all prediction results and the advantages of linear and non-linear models, the best model for predicting tea polyphenol contents will be selected.

5. Conclusions

In this study, NIR spectroscopy and three PLS models were applied to predict tea polyphenols contents in Fuzhuan tea rapidly and non-destructively. Among the three PLS models used, PLS was the worst, followed by si-PLS, whereas the best was ga-PLS, which was established using 37 spectral data points. The optimal calibration model was achieved with Rp2 = 0.9996 and RMSEP = 0.0611% in the prediction set. The P-value between the true and predicted values in the prediction set samples was 1.46 (P > 0.05), showing no significant difference. These experimental results provide a reference for the rapid detection of tea polyphenols contents in Fuzhuan tea.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Natural Science Foundation of Heilongjiang Province for support [LH2020C063] for this research project.

References

Ahmed, R. S. I., Soave, C., Edlbauer, T. G., Patel, K. R., Elghoul, Y., Oliveira, A. V. P. D., & Dou, Q. P. (2019). Discovery of green tea polyphenol-based antitumor drugs. *Mechanisms of Action and Clinical Implications*, 7, 313–332. https://doi.org/10.1007/978-3-030-31269-5_14

Blanco, M., & Peguero, A. (2010). Analysis of pharmaceuticals by NIR spectroscopy without a reference method. *Trends in Analytical Chemistry*, 29(10), 1127–1136. https://doi.org/10.1016/j.trac.2010.07.007

Ceci, C., Lecal, P. M., Tentori, L., Martino, M. G. D., Milano, R., & Graziani, G. (2018). Experimental evidence of the antitumor, immunostimulatory and antiangiogenic activity of elagic acid. *Nutrients*, 10(11), 1756. https://doi.org/10.3390/nu10111756

Chen, G., Bai, Y., Zeng, Z., Peng, Y., Liu, Z., Shen, W., Zeng, X., & Liu, Z. (2021). Structural characterization and immunostimulatory activity of heteropolysaccharides from fuzhuan brick tea. *Journal of Agricultural...
Punyasiri, G., Ghosh, S., Chen, N., NIR https://doi.org/10.1016/j.jnutbio.2021.108634
Ranatunga, R., Journal jafc.0c06913, 2019, 1–16. Recent insights on tea metabolites, their biosynthesis and chemo-preventing effects: A review. Critical Reviews in Food Science and Nutrition, 7, 1–20. https://doi.org/10.1080/10408398.2019.1848471
Sheehan, T. J., Hamnett, H. J., Beasley, R., & Fitzmaurice, P. S. (2019). Chemical and physical variations of cannabis smoke from a variety of cannabis samples in New Zealand. Forensic Sciences Research, 4(2), 168–178. https://doi.org/10.1002/jsfa.9123
Silva, F. E. B. D., Flores, E. M. D. M., Parisseto, G., Müller, E. I., & Ferrão, M. F. (2016). Green method by diffuse reflectance infrared spectroscopy and spectral region selection for the quantification of sulphamethoxazole and trimethoprim in pharmaceutical formulations. Anais da Academia Brasileira de Ciencias, 88(1), 1–15. https://doi.org/10.1590/0100-874520168005075
Tavaanaie, M. A., Esmailnia, N., & Mojtahedi, M. R. M. (2015). Olive hue visible–near infrared camouflaging properties of high speed melt spun poly (Ethylene terephthalate) multilaminate yarn. Dyes and Pigments, 114, 267–272. https://doi.org/10.1016/j.dyepig.2014.11.019
Wang, J., Wang, Y., Cheng, J., Wang, J., Sun, X., Sun, S., & Zhang, Z. (2018). Enhanced cross-category models for predicting the total polyphenols, caffeine and free amino acids contents in Chinese tea using near-infrared spectroscopy. Lwt–Food Science and Technology, 96, 90–97. https://doi.org/10.1016/j.lwt.2018.05.012
Wu, C., Wu, T., Hashmonay, R. A., Chang, S., Wu, Y., Chao, C., Kagan, R. H., Chase, M. J., & Kagan, R. H. (2014). Measurement of fugitive volatile organic compound emissions from a petrochemical tank farm using open-path Fourier transform infrared spectrometry. Atmospheric Environment, 82, 333–342. https://doi.org/10.1016/j.atmosenv.2013.10.036
Xue, L., & Jian, L. (2014). Using short wave visible–near infrared reflectance spectroscopy to predict soil properties and content. Spectroscopy Letters, 47(10), 729–739. https://doi.org/10.1080/00376398.2013.840315
Yang, Y., Wang, L., Wu, Y., Liu, X., Bi, Y., Xiao, W., & Chen, Y. (2017). On-line monitoring of extraction process of Flos Lonicerae Japoniceae using near infrared spectroscopy combined with synergy interval PLS and genetic algorithm. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 182, 73–80. https://doi.org/10.1016/j.saa.2017.04.004
Zhang, L., & Fang, Y. (2018). Research progress of the effect of Fuzhuan tea. Food Research and Development, 76(4), 782–790. https://doi.org/10.1002/jsfa.10353
Zhou, D. D., Saimaiti, A., Luo, M., Huang, S. Y., Xiong, R. G., Shang, A. R., & Li, H. B. (2022). Fermentation with tea residues enhances antioxidant activities and polyphenol contents in koubucha beverages. Antioxidants, 11(11), 1555. https://doi.org/10.3390/antiox11111555
Zhou, M., Tian, X., Wu, Z., Li, K., & Li, Z. (2021). Fuzhuan brick tea supplemented with areca nuts: Effects on serum and gut microbiota in mice. Journal of Food Biochemistry, 45(5), e13737. https://doi.org/10.1111/jfbc.13737
Zhou, X., Yang, Z., Haughey, S. A., Galvin-King, P., Han, L., & Elliott, C. T. (2015). Classification the geographical origin of corn distillers dried grains with solubles by near infrared reflectance spectroscopy combined with chemometrics: A feasibility study. Food Chemistry, 189, 158–168. https://doi.org/10.1016/j.foodchem.2015.05.088
Zhu, Y., Chen, J., Ji, X., Xu, X., Ling, T., Zhang, Z., & Wan, X. (2015). Changes of major tea polyphenols and production of four new B-ring fission metabolites of catechins from post-fermented Jing-Wei Fu brick tea. Food Chemistry, 170, 110–117. https://doi.org/10.1016/j.foodchem.2014.09.104
Zhu, N., Wang, K., Zhang, S. L., Zhao, B., & Wang, S. W. (2020). Application of artificial neural networks to predict multiple quality of dry-cured ham based on protein degradation. Food Chemistry, 344 (12), 128586. https://doi.org/10.1016/j.foodchem.2020.128586