Computers and Electronics in Agriculture 175 (2020) 105538

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Onsite nutritional diagnosis of tea plants using micro near-infrared spectrometer coupled with chemometrics

Yu-Jie Wang, Shan-Shan Jin, Meng-Hui Li, Ying Liu, Lu-Qing Li, Jing-Ming Ning*, Zheng-Zhu Zhang*

State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei 230036, China

ARTICLE INFO

Keywords:
Tea plants
Pigment contents
Micro near-infrared spectrometer
Variable combination population analysis
VCPA-GA hybrid strategy

ABSTRACT

A rapid and accurate diagnosis of nutritional status in field crops is crucial for site-specific fertilizer management. The micro near-infrared spectrometer (Micro-NIRS) is an extremely portable optical device that can be connected to a smartphone through a Bluetooth connection. In this study, a Micro-NIRS was used to evaluate pigment contents, namely chlorophyll *a* (Chl-a), chlorophyll *b* (Chl-b), and carotenoid (Car) in two varieties of field tea plants. A variable combination population analysis (VCPA), genetic algorithm (GA), and VCPA-GA hybrid strategy were used to select characteristic wavelengths; a partial least squares regression (PLSR) algorithm was employed for modeling. Results indicated that the simplified VCPA-GA-PLSR models provided the most favorable performance among all models for Chl-a, Chl-b, and Car content prediction; the correlation coefficients in prediction (Rps) were 0.9226, 0.9006, and 0.8313, respectively; the root mean square errors in prediction (RMSEPs) were 0.0952, 0.0771, and 0.0373 mg/g, respectively; the relative prediction deviations (RPDs) were 2.55, 1.92, and 1.79, respectively. Extracted characteristic variables occupied < 13.63% of full spectra. The current work provided a useful example for implementing a smartphone-based Micro-NIRS system that can diagnose plant nutrition rapidly, nondestructively, and at low cost.

1. Introduction

Tea (*Camellia sinensis* L.) is a commercial crop, which is widely cultivated in China with a cultivation area of 2,930,500 ha in 2018. Fertilizer application, especially nitrogen application, considerably increases the yield and improves the quality of tea plant leaves. However, excessive or ineffective nitrogen application not only increases agricultural costs but also causes pollution of the environment and underground water (Saraswathy et al., 2007). The physiological indexes of each tea plant are closely related to the growth status and yield of tea leaves from that plant. Photosynthetic pigments, including chlorophyll and carotenoids, play essential roles in photosynthesis, and can be used as crucial indicators of plant nutrition status (Nishio, 2000). Studies have indicated that foliar chlorophyll concentrations are positively correlated with nitrogen content, which can help farmers to assess the nitrogen status of a crop and to optimize fertilization (Wang et al., 2019). Carotenoids provide much complementary information on vegetation physiological status and protect chlorophyll molecules from photo-oxidation under excessive light (Ge et al., 2011). The accurate diagnosis of photosynthetic pigment content in tea leaves is conducive for understanding the nutritional status of tea plants and providing rational fertilization.

Currently, many methods are used for quantitatively determining the contents of photosynthetic pigments in plants, such as spectrophotometry and high-performance liquid chromatography (HPLC) (Solymosi et al., 2012). However, these methods are laboratory-based, destructive to samples, time-consuming, and demand technical skills. Moreover, they require complex sample preparation and cannot provide determination. Therefore, accurate, rapid, and nondestructive determination methods must be developed. Near-infrared (NIR) spectroscopy is a mature detection technology with the advantages of rapidity, nondestructiveness, accuracy, and reliability. NIR spectroscopy can be used to describe the overtone and combination bands of C–H, O–H, and N–H groups in the spectral wavelength range of 780–2500 nm for analyzing most chemical compounds (Guo et al., 2016; Cui et al., 2019). Because photosynthetic pigments such as chlorophyll and carotenoids have hydrogen-containing groups, NIR can capture the relevant information of these pigments. Based on these principles, NIR was used by researchers to assess the photosynthetic pigment content of various plants (Zhang et al., 2016). He et al. reviewed the progress of

* Corresponding authors.
E-mail addresses: ningjm1998009@163.com (J.-M. Ning), zzz@ahau.edu.cn (Z.-Z. Zhang).

https://doi.org/10.1016/j.compag.2020.105538
Received 1 March 2020; Received in revised form 27 May 2020; Accepted 27 May 2020

spectral techniques applied to crops for the diagnosis of nutrient status. They concluded that NIR combined with chemometrics can be used to accurately predict the pigment content in various crops (He et al., 2015). However, most of the studies employed laboratory-based benchtop NIR spectrometers, which are costly and immobile. A small number of studies have used portable NIR spectrometers to diagnose pigments, such as ASD field spectrometers (Liu et al., 2019). However, the high cost of these instruments and the need of operation skills make it difficult for them to become routine testing tools for farmers. Therefore, it is necessary and advantageous for farmers to search for low-cost and extremely portable diagnostic tools for plant pigments for scientific crop management.

The micro near-infrared spectrometer (Micro-NIRS) is an extremely portable optical device that can be connected to a smartphone through a Bluetooth connection. It can acquire, record, and store spectral data, then upload data to cloud servers. Due to advantages over lab-based desktop NIRS in portability, price, and environmental requirements, Micro-NIRS systems have received considerable attention from researchers. Malegori et al. assessed the feasibility of one of the smallest NIR spectrometers on the market (MicroNIR 1700) for the evaluation of acerola fruit quality (titratable acidity and ascorbic acid content) during ripening. The results indicated that the predictive ability of the MicroNIR 1700 was comparable with that of a desktop FT-NIR spectrometer (Malegori et al., 2017). Similarly, Sun et al. used a MicroNIR 1700 for determining the glucosamine content in fermentation. The results of Passing–Bablok regression and paired t testing indicated no significant differences between the MicroNIR 1700 and FT-NIR spectrometers (Sun et al., 2018). Coronel-Reyes et al. employed a low-cost Micro-NIRS connected to a smartphone to determine egg storage time. Their success demonstrated that such devices can quantify an egg's freshness to an industrial standard of precision (Coronel-Reyes et al., 2018). These results indicated that the performance of the Micro-NIRS was comparable to that of any desktop NIRS in the qualitative and quantitative analysis of food quality. However, few studies have been conducted on evaluations for the stress diagnosis and nutritional diagnosis of plants.

Thus, the main goals of this study were to investigate the analytical performance of a smartphone-based Micro-NIRS and to evaluate the prediction accuracy in terms of direct applicability in the field. This study explored the effectiveness of a smartphone-based Micro-NIRS for estimation of tea plant photosynthetic pigment content under field conditions. The results of this novel study is expected to help in the onsite, rapid, and low-cost evaluation for the photosynthetic pigment content in tea plants, which will quickly diagnose and evaluate the nutritional status of tea plants. These promising results can provide support for the development of smartphone applications in the near future.

2. Materials and methods

2.1. Sample preparation

Two varieties of tea plant, namely Nongkangzao (NKZ) and Longjing 43 (LJ 43), were selected as research materials. Both varieties have been identified and registered at the national level and have large areas of cultivation in China, where they are often used for green tea. We conducted our trials at High-tech Agricultural Garden in Anhui Agricultural University, Hefei city, Anhui province (31.90° N, 117.23° E).

2.2. Spectral data acquisition

In this study, we used a micro-NIRS (NIR-S-R2; InnoSpectra Corporation, Taiwan, China) and a smartphone (Huawei Honor 10GT; Huawei Technology Co., Ltd., Shenzhen, China) for the onsite spectral data collection of tea leaves (Fig. 1a). The spectrometer was connected to the smartphone via a Bluetooth connection. The Micro-NIRS used exhibited a miniature size with the length, width, and height of 75, 58, and 26.5 mm, respectively, and weighed approximately 77 g. Its low cost and reduced dimensions allow users to perform tests rapidly, which could be developed and integrated, for example, into smartphones. And the smartphone application was provided by the manufacturer (InnoSpectra Corporation, Taiwan, China) of the micro-NIR spectrometer. According to the manufacturer's instructions, the application was downloaded and installed on a smartphone running with Android system, and used only to collect and store spectral data, in the format of *.cvs*. The spectrum files stored on the smartphone were then imported into an external computer via a universal serial bus (USB) line. The transferred data was converted into a matrix format in the computer, which is implemented in Excel software. Finally, the processed data was imported into MATLAB (Mathworks Inc., Natick, MA, USA) in computer to build the prediction models.

To avoid the randomness of different scanning areas, each leaf was scanned at ten points (Fig. 1b), and the averaged spectrum of the resulting ten scans was considered the representative spectral data of the leaf. The spectral acquisition process was performed in an absorption mode, and spectral data were saved as absorbance values (Fig. 1c). Eventually, the spectral data in the range of 900–1700 nm for 72 tea leaf samples (42 NKZ, 30 LJ 43) were acquired.

2.3. Pigment content measurement

After the onsite measurement of NIR spectral data, the tea leaves were plucked and placed in self-sealing plastic bags and then were transported to the laboratory in an environment of 4 °C. The determined contents of photosynthetic pigments were consistent with our previous reports (Wang et al., 2019). The main vein of the leaf was removed, and the remaining portion of the leaf was cut into pieces. Fresh leaf samples weighing 0.1 g were placed in a 10-mL centrifuge tube with 95% ethanol solution, and the leaf was soaked in the solution for 24 h in the dark environment. Subsequently, the centrifuge tube was centrifuged at a speed of 3500 r/min for 10 min before the supernatant was analyzed using a UV spectrophotometer (U-5100, Hitachi Ltd., Tokyo, Japan) at wavelengths of 665, 649, and 470 nm. The calculation equations of pigment content that were used refer to previous reports (Wang et al., 2019).

2.4. Establishment of quantitative prediction model

2.4.1. Sample division

Before establishing a quantitative calibration model, all samples should be divided into the calibration and prediction sets. Here, we employed the sorting method. First, all samples were arranged in ascending order according to the content of pigments. Then, in each group of three samples the middle sample was considered the prediction set and the other two were considered the calibration set. The resulting calibration and prediction sets contained 48 and 24 samples, respectively.

2.4.2. Quantitative model using full spectrum

NIR spectrum contain useful information related to chemical composition and molecular structure that is not directly available from their spectral reflectance values. The quantitative model is to establish the relationship between the spectral information and the target attributes by means of multivariate modeling. Thus, a relational model can be used to predict the target attributes of unknown samples in a rapid and nondestructive way.

In this study, partial least squares regression (PLSR) algorithm was used for development of Chl-a, Chl-b, and Car determination models based on the full spectrum. PLSR is a highly effective multivariate algorithm for quantitative modeling; it is particularly advantageous for problems large sets of collinear variables (Wold et al., 2001). PLSR is a

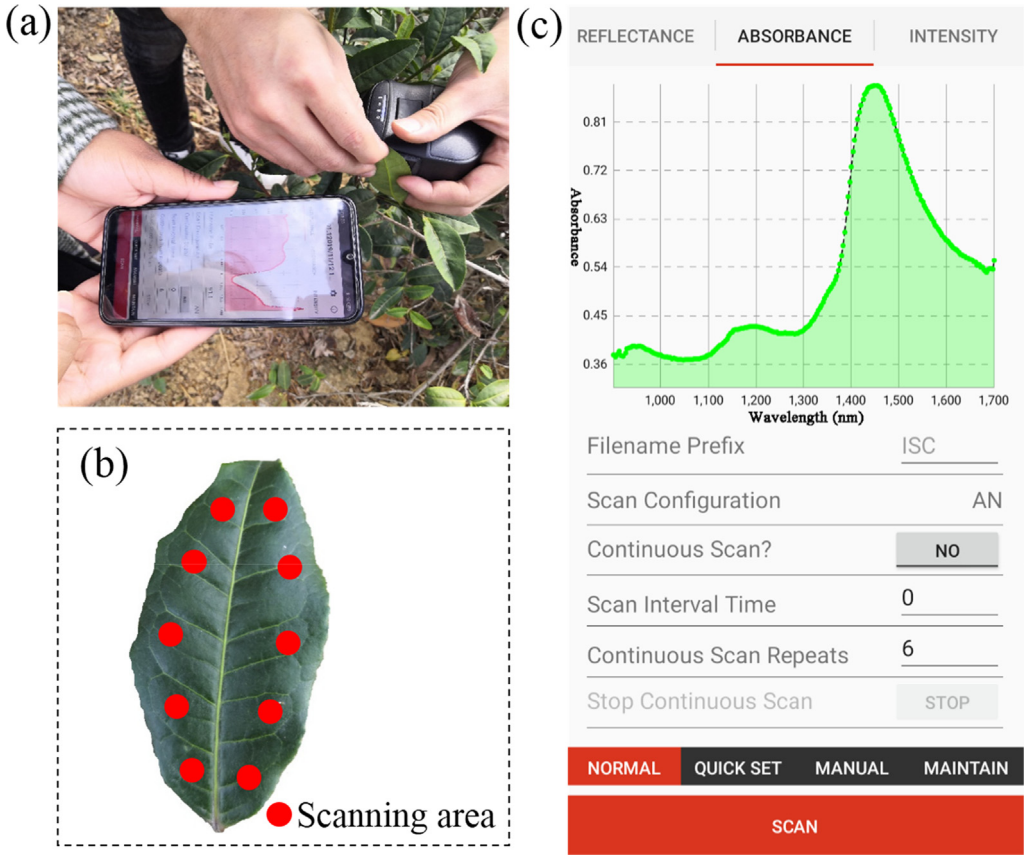

Fig. 1. Photo of smartphone-based Micro-NIRS system for on-site measurement in tea garden (a); ten different scanning areas in one tea leaf sample (b); and spectrum acquisition interface on smartphone (c).

linear regression algorithm which simultaneously projects the independent spectral variables (X) and dependent target variable (Y) to a new space with the constraint that the components could explain as much as possible of the variance between X and Y. It compresses the spectral data into a set of orthogonal variables called latent variables (LVs), which carry the most information and maximum covariance between the X and Y (Barbin et al., 2012; Huang et al., 2017). And the optimal number of LVs were determined by minimizing the error sum of squares, which is normally achieved by cross-validation. Since few LVs used for modeling will lead to the loss of useful information in spectral variables, while too many LVs will increase the dimension and useless information. The maximum number of LVs is usually set to 10–20. And in this study, we set the maximum number of LVs to 15 for preventing the model from overfitting due to many LVs.

2.4.3. Quantitative model using characteristic wavelengths

The acquired full spectral information contained redundancy and collinearity, which led to complex and inefficient modeling processes. One solution for this problem was selecting the characteristic wavelengths (CWs) from the full spectrum and removing extraneous spectral variables. Then, these CWs-based PLSR prediction models were established. CWs are crucial wavelengths that carry the most useful information regarding the pigment content. A variable combination population analysis (VCPA), genetic algorithm (GA), and VCPA combined with GA (VCPA-GA) were used to select CWs.

VCPA is a novel and effective algorithm proposed by Yun et al. (2015). In VCPA, the exponentially decreasing function (EDF), an effective principle of 'survival of the fittest' from Darwin's natural evolution theory, is used to determine the number of variables to keep and continuously shrink the variable space. And in each EDF run, a binary matrix sampling (BMS) strategy is employed to give each spectral variable the same chance to be selected and generate different spectral variable combinations. BMS is used to produce a population of subsets to construct a population of sub-models. Then, model population analysis (MPA) is employed to find the spectral variable subsets with the lower root mean squares error of cross validation (RMSECV). The frequency of each spectral variable appearing in the best 10% sub-models is computed. The higher the frequency is, the more important the spectral variable is. And the detailed flow chart of VCPA consists of the following steps:

Step 1: Set running parameters: the occupancy percentage value of '1'per column as α, the number of sampling runs as k, the EDF runs as N.

Step 2: When $i \leq N$ runs of EDF, generate a binary matrix $\mathbf{M}$ ($k \times p$) and randomly permutation by column. The number of '1' in each column is $k\alpha$.

Step 3: While $j \leq k$, calculate the RMSECV value of the subset obtained from jth row of $\mathbf{M}$. And sort all the k RMSECV and obtain the subsets that are the best $\sigma = 10\%$ sub-models with the lowest RMSECV value.

Step 4: Computer the ratio of variables to be remained using EDF: $r_i = e^{-\theta i}$, and retain $p = p \times r_i$ variables based on the frequency. Then get a new p.

Step 5: After N runs, there are $\omega = 14$ variables left. Calculate the RMSECV of all combinations among 14 variables, and record the lowest one as BEST.

Step 6: Finally, choose the subset as optimal subset from BEST.

In our study, all of the key parameters were the default values in the downloaded code. The EDF runs (N) was set as 50, the BMS sampling runs (k) was 1000, and the number of the left variables in the final run of EDF (ω) was 14. The values for these parameters are optimized according to Yun et al. (2015).

Table 1
Statistical analysis of measured photosynthetic pigments content (mg/g, fresh weight) of in the canopy leaves of different tea varieties.

Variety	Chl-a content			Chl-b content			Car content		
	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD
NKZ	1.79–2.36	2.15	0.13	0.66–1.04	0.87	0.09	0.28–0.65	0.47	0.07
LJ 43	1.67–1.92	1.66	0.17	0.44–0.80	0.64	0.09	0.31–0.52	0.44	0.05
Total	1.67–2.36	1.88	0.25	0.44–1.04	0.77	0.15	0.28–0.65	0.46	0.07

NKZ: Nongkangzao; LJ 43: Longjing 43; Chl-a: chlorophyll *a*; Chl-b: chlorophyll *b*; Car: carotenoids; SD: standard deviation.

GAs have been successfully applied for spectral variable selection problems in many works. GAs were originally inspired by the theory of biological evolution and natural selection, in which variables that produce a high-performing calibration model have a high probability of being retained and included in the selected variable set in subsequent model calibration. GA-PLS models have been used for the optimization of various linear prediction models. In our study, the parameters of the GA were set according to Yun et al. (2015) as follows: the probability of crossover was set as 50%, the probability of mutation was set as 1%, and the number of runs was set as 100.

VCPA-GA is a VCPA-based hybrid strategy proposed by Yun et al. (2019). It continuously shrinks the variable space from big to small and optimizes it based on modified VCPA in the first step. It then employs a GA to carry out further optimization in the second step. It uses all the effective aspects of VCPA and GA, and if numerous variables are required, VCPA-GA outperforms both VCPA and GA. The VCPA-GA strategy consists of the following two steps:

Step 1: VCPA is conducted to shrink the variable space. For VCPA, in this work, the number of variables left in the *N*th run is set to 100 in the EDF step, which leaves 100 variables for further optimization by GA.

Step 2: GA optimizes the remaining 100 variables over *N* iterations. Subsets of these 100 variables are retained; other variables with little contribution are eliminated by EDF. The space spanned by the remaining variables is small and optimized, making it easier for GA to select the optimal variable subset.

More details about VCPA, GA, and VCPA-GA can be found in previous works (Yun et al., 2015, 2019).

2.4.4. Model evaluation

The performance levels of calibration models, namely PLSR, VCPA-PLSR, GA-PLSR, and VCPA-GA-PLSR, were evaluated by the correlation coefficient of calibration (Rc, Eq. (1)), and root mean square errors in calibration (RMSEC, Eq. (2)). The predictive precision was evaluated using the correlation coefficient of prediction (Rp, Eq. (3)) and root mean square error in prediction (RMSEP, Eq. (4)). The relative prediction deviation (RPD, Eq. (5)), defined as the ratio of the standard deviation in the prediction set to the RMSEP, has been previously used in model evaluation. An accurate model should have high values of Rc, Rp, and RPD, and low RMSEC and RMSEP values.

$$\mathrm{Rc} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_{cal} - y_{act})}{\sum_{i=1}^{n} (y_{cal} - y_{mean})}} \quad (1)$$

$$\mathrm{RMSEC} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{cal} - y_{act})^2} \quad (2)$$

$$\mathrm{Rp} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_{pre} - y_{act})}{\sum_{i=1}^{n} (y_{pre} - y_{mean})}} \quad (3)$$

$$\mathrm{RMSEP} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{pre} - y_{act})^2} \quad (4)$$

$$\mathrm{RPD} = \frac{SD}{RMSEP} \quad (5)$$

where, n is the number of samples used in model development; y_{cal} is the predicted value in calibration set; y_{act} is the actual value measured by chemical method; y_{mean} is the average value; y_{pred} is the predicted value in prediction set; and SD is the standard deviation of measured values in prediction set.

2.5. Software

The procedures of PLSR modeling, wavelength selection by VCPA, GA, and VCPA-GA were implemented in MATLAB R2014a (Mathworks Inc., Natick, MA, USA).

3. Results and discussion

3.1. Pigment content distribution in the canopy leaves of the two tea plant varieties

Photosynthetic pigments, including chlorophyll and carotenoids, are crucial nutrient components of tea plants and can indicate their growth status. Moreover, chlorophyll is an essential green-color related component of green tea, which is positively correlated with the quality of green tea. We conducted statistical analysis of photosynthetic pigment content in the canopy leaves of two commercially noteworthy varieties of green tea, namely NKZ and LJ 43. As listed in Table 1, the Chl-a values in NKZ and LJ 43 varieties were 2.15 ± 0.13 (mean value ± standard deviation) and 1.66 ± 0.17 mg/g, respectively. The values of Chl-b in NKZ and LJ 43 varieties were 0.87 ± 0.09 and 0.64 ± 0.09 mg/g, respectively. These results illustrated that the chlorophyll levels in NKZ leaves were higher than those in LJ 43. However, the levels of Car in NKZ and LJ 43 varieties were 0.47 ± 0.07 and 0.44 ± 0.05 mg/g, respectively, indicating negligible differences between two varieties.

The results of the division of calibration and prediction sets are presented in Table 2. For both chlorophyll and carotenoid contents, the content span of the calibration set was larger than that of the prediction set, indicating that the prediction set can be used to verify the performance of the established model and the division of the sample set were reasonably appropriate.

3.2. Spectral curve characteristics produced by Micro-NIRS

The spectral curves of 72 tea leaf samples in the range of 900–1700 nm are illustrated in Fig. 2. Fig. 2 indicates that the spectral curve trends of all samples were similar. Two obvious absorption peaks

Table 2
Divisions of calibration and prediction sets for pigment content estimation.

Pigment	Dataset	Number	Range	Mean	SD
Chl-a	Calibration set	48	1.25–2.36	1.88	0.25
	Prediction set	24	1.36–2.27	1.88	0.24
Chl-b	Calibration set	48	0.44–1.04	0.78	0.15
	Prediction set	24	0.48–1.03	0.78	0.15
Car	Calibration set	48	0.28–0.65	0.46	0.07
	Prediction set	24	0.31–0.61	0.46	0.07

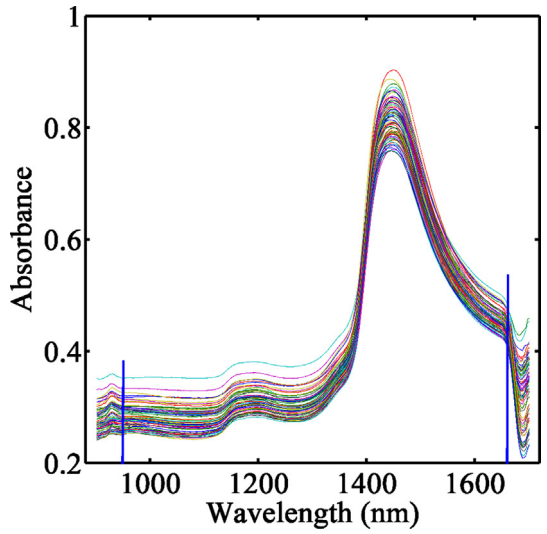

Fig. 2. Raw spectral curves of tea leaf samples obtained using Micro-NIRS.

were observed at approximately 1180 and 1440 nm. A flat absorption peak at approximately 1180 nm may have been caused by the C–H stretching vibration of CH_3 and the first overtone of O–H stretching (Wei et al., 2019; Özdemir et al., 2018), and the absorption peak at approximately 1440 nm was contributed to the first overtone of O–H stretching (Lee et al., 2014). The first absorption peak was related to the high moisture content of a tea leaf, which accounted for approximately 70% of the total weight of the leaf. Furthermore, all spectral curves fluctuated sharply within the spectral ranges of 900–950 and 1650–1700 nm, indicating that noise and interferences existed in these parts. Therefore, only spectral data in the spectral range of 950–1650 nm were used for further analysis and modeling.

3.3. Pigment content prediction by using full-PLSR models

A PLSR algorithm was employed to develop cross-variety models for predicting the Chl-a, Chl-b, and Car contents of the two tea plant varieties. Moreover, the optimal number of LVs was determined using the lowest RMSECV value. The pigment content prediction of full-PLSR models is presented in Table 3. For Chl-a, the PLSR model that employed full spectra provided a favorable performance with 5 LVs and the Rc and RMSEC of 0.9548 and 0.0739 mg/g, respectively. However, in the prediction set, the correlation coefficient decreased to 0.9084 and the RMSEP increased to 0.1050 mg/g, indicating that the performance of the model was unstable. The RPD value of the model was 2.31, and the model exhibited a high accuracy and wide applicability (Lohr et al., 2017). For Chl-b, the PLSR model provided the Rc, RMSEC, Rp, and RMSEP of 0.9243, 0.0566 mg/g, 0.8509, and 0.0946 mg/g, and an RPD value of 1.57 with 11 LVs. For Car content, the Rc, RMSEC, Rp, and RMSEP values were 0.8059, 0.0402 mg/g, 0.7934, and 0.0414 mg/g, respectively, and the RPD value was 1.61 with 7 LVs. According to the study of Saeys et al. (2005), the scheme of RPD values for inhomogeneous samples was used as follows: < 1.5, unusable; approximately 1.5–2.0, distinguishing between low and high samples; approximately 2.0–2.5, semiquantitative rating; > 2.5, quantitative rating. In this study, full-PLSR models for Chl-a, Chl-b, and Car content prediction provided different performances: the PLSR model performed more favorably for Chl-a than for Chl-b and Car, possibly because the content of Chl-a was higher than that of Chl-b and Car; therefore, in the near infrared region, the spectral response intensity of Chl-a was higher than that of Chl-b and Car. By contrast, many variables irrelevant to the target compound content in the full spectrum may have existed, which reduced the accuracy of the calibration model. Thus, to improve the performance of the prediction model, we employed different variable selection methods to seek the most crucial CWs.

Table 3
Results of PLSR models using full spectrum for pigment content prediction.

Pigment	Parameter	Calibration set		Prediction set		
		Rc	RMSEC	Rp	RMSEP	RPD
Chl-a	LVs = 5	0.9548	0.0739	0.9084	0.1050	2.31
Chl-b	LVs = 11	0.9243	0.0566	0.8509	0.0946	1.57
Car	LVs = 7	0.8059	0.0402	0.7934	0.0414	1.61

LVs: number of latent variables in PLSR models; Rc: correlation coefficient of calibration; RMSEC: root mean square errors in calibration; Rp: correlation coefficient of prediction; RMSEP: root mean square errors in prediction; RPD: residual predictive deviation.

3.4. Pigment content prediction by using CW-PLSR models

Three CW selection methods, namely VCPA, GA, and VCPA-GA, were used to develop CW-PLSR models. However, the sampling strategies of the three algorithms were random. Therefore, the three algorithms were run 50 times each to evaluate the reproducibility and stability of obtained results.

3.4.1. VCPA-PLSR models

CWs selected using VCPA method for predicting Chl-a, Chl-b, and Car are presented in Table 4. Especially, 13, 10, and 8 CWs were selected from the full 198 spectral wavelengths, indicating that irrelevant variables accounting for a minimum of 93.43% of all spectral variables were eliminated. On the basis of these selected CWs, the VCPA-PLSR models were separately established, and results are presented in Table 5. For predicting Chl-a, Chl-b, and Car, VCPA-PLSR models provided the Rp of 0.8722, 0.8764, and 0.7759, respectively; the RMSEP of 0.1260, 0.0832, and 0.0463 mg/g, respectively; and the RPD of 1.93, 1.78, and 1.44, respectively. These results indicated that the accuracies of the prediction model for Chl-a and Chl-b were acceptable; however for Car the model performed inadequately.

3.4.2. GA-PLSR models

The GA is a widely used variable selection algorithm and is often used in combination with PLSR. The GA selected 12, 15, and 34 spectral variables for Chl-a, Chl-b, and Car, respectively (Table 4). The performance of GA-PLSR models is provided in Table 5. For the prediction of Chl-a, Chl-b, and Car the Rp values were 0.9144, 0.8583, and 0.7804, respectively; the RMSEP values were 0.0991, 0.0898, and 0.0437 mg/g, respectively; and the RPD values were 2.45, 1.65, and 1.53, respectively. In terms of the evaluation index of GA-PLSR models, the model for Chl-a achieved a high prediction accuracy, and the prediction models for Chl-b and Car were acceptable because their RPD values exceeded 1.5.

3.4.3. VCPA-GA-PLSR models

Considering the advantages of the VCPA and GA, Yun et al. proposed a hybrid strategy of the VCPA-GA to select CWs (Yun et al., 2019). CWs related to pigment content through the VCPA-GA algorithm are presented in Table 4. For Chl-a, Chl-b, and Car seven, eight, and twenty-seven CWs were selected, respectively, indicating that irrelevant variables accounting for a minimum of 86.36% of all spectral variables were eliminated. Subsequently, VCPA-GA-PLSR models were separately established for predicting Chl-a, Chl-b, and Car (Table 5). For Chl-a, Chl-b, and Car, VCPA-GA-PLSR models provided Rp of 0.9226, 0.9006, and 0.8313, respectively; RMSEP of 0.0952, 0.0771, and 0.0373 mg/g, respectively; and RPD of 2.55, 1.92, and 1.79, respectively. In particular, for Chl-a prediction, the RPD value exceeded 2.5, indicating that the VCPA-GA-PLSR model for Chl-a was satisfactory. Moreover, VCPA-GA-PLSR models for Chl-b and Car were acceptable, because their RPD

Table 4
Selected characteristic wavelengths by VCPA, GA, and VCPA-GA methods for pigment content prediction.

Pigment	Method	Number	Wavelength (nm)
Chl-a	VCPA	13	953.17, 985.16, 992.79, 1201.94, 1206.70, 1293.32, 1307.12, 1310.56, 1318.57, 1325.42, 1335.66, 1452.27, 1525.01
	GA	12	985.16, 992.79, 996.60, 1184.04, 1307.12, 1310.56, 1325.42, 1360.52, 1405.15, 1449.02, 1485.71, 1518.68
	VCPA-GA	7	964.72, 985.16, 1176.84, 1184.04, 1360.52, 1405.15, 1426.08
Chl-b	VCPA	10	1173.24, 1180.44, 1201.94, 1206.70, 1445.75, 1455.53, 1515.51, 1538.67, 1599.64, 1634.02
	GA	15	1173.24, 1180.44, 1198.37, 1206.70, 1322.00, 1373.99, 1439.21, 1455.53, 1469.59, 1499.59, 1502.78, 1515.51, 1599.64, 1634.02, 1646.03
	VCPA-GA	8	1173.24, 1180.44, 1445.75, 1455.53, 1496.39, 1532.37, 1548.07, 1608.79
Car	VCPA	8	996.60, 1024.41, 1210.26, 1509.15, 1544.94, 1570.91, 1623.97, 1646.03
	GA	34	973.68, 1048.24, 1063.20, 1066.93, 1074.38, 1078.09, 1081.81, 1097.85, 1101.54, 1180.44, 1191.21, 1194.79, 1198.37, 1210.26, 1220.92, 1367.27, 1485.71, 1488.92, 1502.78, 1509.15, 1512.33, 1544.94, 1567.81, 1570.91, 1583.26, 1586.34, 1592.49, 1608.79, 1617.91, 1623.97, 1640.03, 1643.03, 1646.03, 1649.02
	VCPA-GA	27	981.33, 996.60, 1024.41, 1048.24, 1066.93, 1081.81, 1101.54, 1184.04, 1201.94, 1210.26, 1426.08, 1445.75, 1449.02, 1463.11, 1485.71, 1509.15, 1532.37, 1544.94, 1570.91, 1583.26, 1586.34, 1605.74, 1623.97, 1637.03, 1643.03, 1646.03, 1649.02

VCPA: variable combination population analysis; GA: genetic algorithm.

Table 5
Results of PLSR models using characteristic wavelengths selected by VCPA, GA, and VCPA-GA for pigment content prediction.

Pigment	Model	LVs	Calibration set		Prediction set		
			Rc	RMSEC	Rp	RMSEP	RPD
Chl-a	VCPA-PLSR	13	0.9796	0.0499	0.8722	0.1260	1.93
	GA-PLSR	9	0.9699	0.0607	0.9144	0.0991	2.45
	VCPA-GA-PLSR	3	0.9583	0.0709	0.9226	0.0952	2.55
Chl-b	VCPA-PLSR	10	0.9612	0.0403	0.8764	0.0832	1.78
	GA-PLSR	8	0.9562	0.0427	0.8583	0.0898	1.65
	VCPA-GA-PLSR	6	0.9682	0.0366	0.9006	0.0771	1.92
Car	VCPA-PLSR	6	0.9045	0.0286	0.7759	0.0463	1.44
	GA-PLSR	6	0.8631	0.0339	0.7804	0.0437	1.53
	VCPA-GA-PLSR	9	0.9016	0.0292	0.8313	0.0373	1.79

values were > 1.5.

3.5. Model comparison and discussion

We comparatively applied full-PLSR and three CW-based PLSR models to achieve accurate model performance and their results are presented in Tables 4 and 5. For Chl-a, the accuracy of VCPA-GA-PLSR model was higher than that of VCPA-PLSR and GA-PLSR models, with low RMSEP values and high Rp and RPD values. Similar to full-PLSR model, VCPA-GA-PLSR model achieved favorable performance. Only seven CWs were involved in modeling, which not only substantially simplified the modeling process but also increased the RPD value of the model by 10.39%. The selected wavelengths were 964.72, 985.16, 1176.84, 1184.04, 1360.52, 1405.15, and 1426.08 nm. For Chl-b, all simplified PLSR models improved the performance compared with the full-PLSR model did, with lower RMSEP values and higher Rp and RPD values, indicating that many irrelevant variables existed in the full spectrum without contributing to the content of Chl-b, which negatively influenced the results of the prediction model. Irrelevant variables were eliminated to varying degrees through effective wavelength selection. VCPA-GA-PLSR provided the highest accuracy with the Rp of 0.9006, RMSEP of 0.0771, and RPD of 1.92. For Car, the performance of the full-PLSR model exceeded that of VCPA-PLSR and GA-PLSR models, with a higher RPD value and a lower RMSEP value, indicating that both the VCPA and GA algorithms are superfluous for improving the model. However, when the VCPA and GA were used together, the accuracy of the VCPA-GA-PLSR model improved beyond that of the full-PLSR model, with a high RPD of 1.79. However, after the two-step screening of the VCPA and GA, only 27 variables were retained and remaining 171 irrelevant variables were removed. Optimal PLSR models for Chl-a, Chl-b, and Car prediction are illustrated in Fig. 3a, 3b, and 3c, respectively. Fig. 3 illustrates that all sample points were scattered on both sides of the oblique line y = x, indicating that predicted values were close to actual values and indicating the accuracy of the optimal models. VCPA-GA-PLSR models provided more accurate values than did full-PLSR and other two simplified PLSR models for Chl-a, Chl-b, and Car. By contrast, the VCPA-GA algorithm demonstrated advantage in data reduction and variable selection by reducing original 198 variables to 27 variables, which considerably simplified the modeling process and improved the model performance. By contrast, the hybrid algorithm exhibited surprising advantages over the use of the VCPA or GA alone. The possible reasons for this finding were as follows: the VCPA-GA hybrid strategy continuously reduced the variable space and optimized it by using VCPA in the first step, and then employed the GA to further optimize the space in the second step. It successfully applied all advantages of VCPA and GA, and yet avoided their disadvantage when operating with high numbers of variables (Yun et al., 2019). Of all the models tested in this study, VCPA-GA-PLS models were able to achieve the most accurate evaluation of pigment content on the spectra; small sets of variables obtained excellent generalization performance.

Non-destructive testing techniques have been widely used for detecting and evaluating nutrition components in plants, including photosynthetic pigments (Zhao et al., 2016; Wang et al., 2019). However, most studies have used hyperspectral imaging or FT-NIR devices, which are expensive and difficult to conduct in the field. Through characteristic variables selection and PLSR linear modeling, the study indicated that accurately predicting the content of Chl-a in the tea leaves of different varieties was feasible. However, the prediction of Chl-b and Car contents can considerably be improved and optimized.

The innovativeness of the current work is listed as the following reasons: 1) We firstly explored the feasibility of micro-NIR spectroscopy connected with smartphones in the diagnosis of pigment in the tea leaves of different varieties. It provides a new and low-cost tool for the acquisition of spectral information and the field diagnosis of plant phenotypes. 2) A novel VCPA combined with GA algorithm was used to select the characteristic wavelengths associated with the pigments in tea leaves. The simplified VCPA-GA-PLSR prediction models were proved to be optimal and can accurately predict chlorophyll content in the tea leaves of different varieties. These promising results can provide support for the development of smartphone applications in the near future.

4. Conclusion

A plant's level of photosynthetic pigment can reflect the physiological status of that plant, and can provide information to guide rational fertilization. In this study, a smartphone-based Micro-NIRS was used for the on-site evaluation of photosynthetic pigment content in tea plants. Full-PLSR models were established for Chl-a, Chl-b, and Car. Moreover, VCPA, GA, and VCPA-GA algorithms were employed to select the

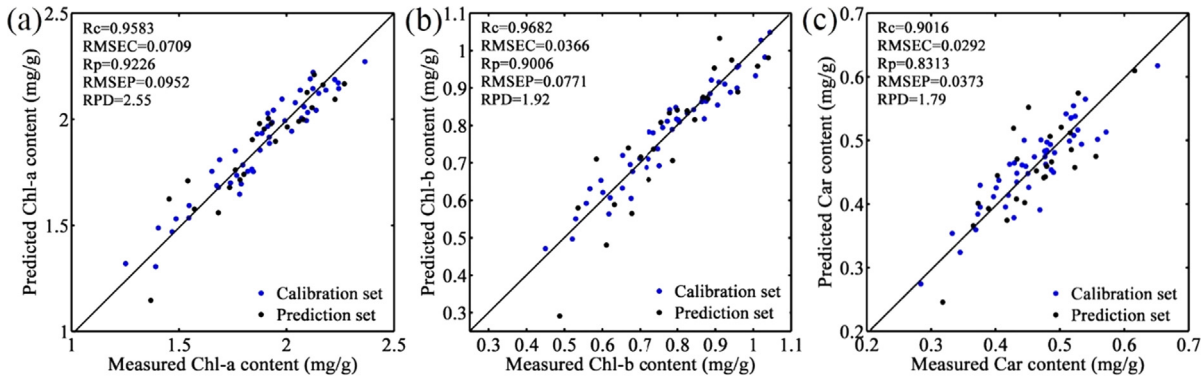

Fig. 3. Scatter plots of optimal VCPA-GA-PLSR models for the prediction of Chl-a (a); Chl-b (b); and Car (c).

characteristic wavelengths. Results showed that the simplified VCPA-GA-PLSR models delivered the best performance for both Chl-a, Chl-b, and Car content prediction, with Rp of 0.9226, 0.9006, 0.8313, RMSEP of 0.0952, 0.0771, 0.0373, and RPD of 2.55, 1.92, and 1.79, respectively. The current work provides a useful example for implementing smartphone-based Micro-NIRS with powerful chemometrics for the diagnosis of plant nutrition; the proposed system is rapid, non-destructive, and affordable. To improve the accuracy and generalization ability of the models, future studies will address tea plant samples of different varieties, growth periods, and nutrient statuses.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yu-Jie Wang: Conceptualization, Data curation, Formal analysis, Methodology. **Shan-Shan Jin:** Formal analysis, Methodology, Software. **Meng-Hui Li:** Data curation, Formal analysis, Methodology. **Ying Liu:** Data curation, Formal analysis. **Lu-Qing Li:** Visualization, Investigation, Supervision. **Jing-Ming Ning:** Writing - review & editing, Validation, Supervision. **Zheng-Zhu Zhang:** Funding acquisition, Project administration, Resources.

Acknowledgements

This work has been financially supported by the National Key Research and Development Program of China (2016YFD0200900, 2017YFD0400800), and Major scientific and technological projects of Anhui Province (18030701149, 18030701153).

References

Barbin, D.F., Elmasry, G., Sun, D.W., Allen, P., 2012. Predicting quality and sensory attributes of pork using near-infrared hyperspectral imaging. Anal. Chim. Acta 719, 30–42.

Coronel-Reyes, J., Ramirez-Morales, I., Fernandez-Blanco, E., Rivero, D., Pazos, A., 2018. Determination of egg storage time at room temperature using a low-cost NIR spectrometer and machine learning techniques. Comput. Electron. Agric. 145, 1–10.

Cui, Y.J., Ge, W.Z., Li, J., Zhang, J.W., An, D., Wei, Y.G., 2019. Screening of maize haploid kernels based on near infrared spectroscopy quantitative analysis. Comput. Electron. Agric. 158, 358–368.

Ge, Z.M., Zhou, X., Kellomaki, S., Wang, K.Y., Peltola, H., Martikainen, P.J., 2011. Responses of leaf photosynthesis, pigments and chlorophyll fluorescence within canopy position in a boreal grass (*Phalaris arundinacea* L.) to elevated temperature and CO2 under varying water regimes. Photosynthetica 49, 172–184.

Guo, Z., Huang, W., Peng, Y., Chen, Q., Ouyang, Q., Zhao, J., 2016. Color compensation and comparison of shortwave near infrared and long wave near infrared spectroscopy for determination of soluble solids content of 'Fuji' apple. Postharvest Biol. Technol. 115, 81–90.

He, Y., Peng, J.Y., Liu, F., Zhang, C., Kong, W.W., 2015. Critical review of fast detection of crop nutrient and physiological information with spectral and imaging technology. Trans. Chin. Soc. Agric. Eng. 31, 174–189.

Huang, H., Shen, Y., Guo, Y.L., Yang, P., Wang, H.Z., Zhan, S.Y., Liu, H.B., Song, H., He, Y., 2017. Characterization of moisture content in dehydrated scallops using spectral images. J. Food Eng. 205, 47–55.

Lee, M.S., Hwang, Y.S., Lee, J., Choung, M.G., 2014. The characterization of caffeine and nine individual catechins in the leaves of green tea (*Camellia sinensis* L.) by near-infrared reflectance spectroscopy. Food Chem. 158, 351–357.

Liu, J.B., Han, J.C., Chen, X., Shi, L., Zhang, L., 2019. Nondestructive detection of rape leaf chlorophyll level based on Vis-NIR spectroscopy. Spectrochim. Acta A Mol. Biomol. Spectrosc. 222, 117202.

Lohr, D., Tillmann, P., Druege, U., Zerche, S., Rath, T., Meinken, E., 2017. Non-destructive determination of carbohydrate reserves in leaves of ornamental cuttings by near-infrared spectroscopy (NIRS) as a key indicator for quality assessments. Biosyst. Eng. 158, 51–63.

Malegori, C., Marques, E.J.N., de Freitas, S.T., Pimentel, M.F., Pasquini, C., Casiraghi, E., 2017. Comparing the analytical performances of Micro-Nir and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms. Talanta 165, 112–116.

Nishio, J.N., 2000. Why are higher plants green? Evolution of the higher plant photosynthetic pigment complement. Plant Cell Environ. 23, 539–548.

Özdemir, I.S., Orturk, B., Celik, B., Santepe, Y., Aksoy, H., 2018. Rapid, simultaneous and non-destructive assessment of the moisture, water activity, firmness and SO2 content of the intact sulphured-dried apricots using FT-NIRS and chemometrics. Talanta 186, 467–472.

Saeys, W., Mouazen, A.M., Ramon, H., 2005. Potential of onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. Biosyst. Eng. 91, 393–402.

Saraswathy, R., Suganya, S., Singaram, P., 2007. Environmental impact of nitrogen fertilization in tea eco-system. J. Environ. Biol. 28, 779–788.

Solymosi, K., Morandi, D., Boka, K., Boddi, B., Schoefs, B., 2012. High biological variability of plastids, photosynthetic pigments and pigment forms of leaf primordia in buds. Planta 235, 1035–1049.

Sun, Z.Y., Li, C., Li, L., Nie, L., Dong, Q., Li, D.Y., Gao, L.L., Zang, H.C., 2018. Study on feasibility of determination of glucosamine content of fermentation process using a micro NIR spectrometer. Spectrochim. Acta A 201, 153–160.

Wang, Y.J., Hu, X., Jin, G., Hou, Z.W., Ning, J.M., Zhang, Z.Z., 2019. Rapid prediction of chlorophylls and carotenoids contents in tea leaves under different levels of nitrogen application based on hyperspectral imaging. J. Sci. Food Agric. 99, 1997–2004.

Wei, Y.Z., Wu, F.Y., Xu, J., Sha, J.J., Zhao, Z.F., He, Y., Li, X.L., 2019. Visual detection of the moisture content of tea leaves with hyperspectral imaging technology. J. Food Eng. 248, 89–96.

Wold, S., Sjostrom, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130.

Yun, Y.H., Bin, J., Liu, D.L., Xu, L., Yan, T.L., Cao, D.S., Xu, Q.S., 2019. A hybrid variable selection strategy based on continuous shrinkage of variable space in multivariate calibration. Anal. Chim. Acta 1058, 58–69.

Yun, Y.H., Wang, W.T., Deng, B.C., Lai, G.B., Liu, X.B., Ren, D.B., Liang, Y.Z., Fan, W., Xu, Q.S., 2015. Using variable combination population analysis for variable selection in multivariate calibration. Anal. Chim. Acta 862, 14–23.

Zhang, J.F., Han, W.T., Huang, L.W., Zhang, Z.Y., Ma, Y.M., Hu, Y.M., 2016. Leaf chlorophyll content estimation of winter wheat based on visible and near-infrared sensors. Sensors 16, 437.

Zhao, Y.R., Li, X.L., Yu, K.Q., Cheng, F., He, Y., 2016. Hyperspectral imaging for determining pigment contents in cucumber leaves in response to angular leaf spot disease. Sci. Rep. 6, 27790. http://data.stats.gov.cn/easyquery.htm?cn=C01&zb=A0D0J&sj=2018.