Prediction Model of Nicotine and Glycerol in Reconstituted Tobacco Leaves Based on Support Vector Machine Algorithm

Qifu Yang, [a] *Ma Kun,*[a] *Liang Ying,*[a] *Wu Jiaquan,*[a] *Zhang Xinyu*,*[a] *and Shuangyan Yang*,*[b]

[a]Kunming University of Science and Technology, 650500 Kunming, China

[b]Yunan Tobacco Biological Technology Co., Ltd., 650500 Kunming, China

Nicotine and glycerol are the two important indexes of reconstituted tobacco and they determine the quality of the reconstituted tobacco. A hand-held near infrared spectrometer was used to collect the spectral data of reconstituted tobacco leaves, and three algorithms of principal component regression, partial least squares and support vector machine were used to build the prediction model of the nicotine and glycerol content in reconstituted tobacco leaves. The experimental results showed that the support vector machine algorithm could achieve the best prediction results compared with principal component regression and partial least squares algorithms. The proposed method can rapidly determine the nicotine and glycerol content of the reconstituted tobacco leaves and it provides a new technical reference for improving the quality of new tobacco.

Keywords: near infrared spectroscopy, reconstituted tobacco leaves, nicotine, propylene glycol, support vector machine

Introduction

As people pay more and more attention to their own health, heating without burning tobacco products rapidly occupy the market of traditional cigarettes due to their advantages of lower content of most harmful components than traditional cigarettes and closest smoking sensation to traditional cigarettes.[1] Reconstituted tobacco was commonly used in tobacco industry, which can effectively control the composition and cost in tobacco processing. Reconstituted tobacco leaves were flake or filament-like regenerated products made from waste tobacco materials generated in the process of cigarette rolling, such as cigarette ends, stems and broken tobacco flakes, and used as cigarette filling materials.[2] There were many indexes to evaluate the quality of reconstituted tobacco. Nicotine (NIC) and glycerol (VG) are two important indexes of reconstituted tobacco.[3] VG can make tobacco leaves have smoother taste. In view of health, remanufactured tobacco requires appropriate NIC and VG content, and the suction sensation should not be blindly pursued while the impact on the body should be ignored.

The determination of NIC and VG components in reconstituted tobacco mainly include chemical analysis method and spectral detection method. Chemical analysis methods were continuous flow method and gas chromatography method, which are expensive and time-comsuming. Near infrared (NIR) spectroscopy is one of the spectral detection methods and the wavelength range is 780-2526 nm. In recent years, many researchers have tried to use NIR spectroscopy to detect the relevant indexes of tobacco leaves, such as NIC, total sugar, reducing sugar, total nitrogen, chlorine,[4] tar, carbon monoxide,[5] ash, total volatile acid, total volatile alkali,[6] potassium[7] and chlorine in tobacco.[8,9] In most of these studies, NIR detection technology combined with partial least square (PLS) method was used to analyze the chemical composition of a tobacco leaf. There was no relevant research on the determination of NIC and VG in reconstituted tobacco leaves. Besides, operators are required to have certain professional knowledge.

In this study, a hand-held near infrared spectrometer was used to collect the spectrum of reconstituted tobacco leaves. Different modeling methods were used to build the NIC and VG models, and the best prediction model was obtained. The experimental results showed that it was feasible to predict the content of NIC and VG in reconstituted tobacco leaves by using hand-held near infrared spectroscopy.

*e-mail: zhangxinyu7212@sohu.com; ysyynbg@163.com
Editor handled this article: Ivo M. Raimundo Jr. (Associate)

Experimental

Equipment and samples

A hand-held NIR spectrometer used in the experiment was the MicroNIR 1700 device provided by Viavi Solution, Milpitas, CA, United States. The parameters of this hand-held NIR spectrometer were set in Table 1.

Table 1. Related parameter settings of MicroNIR

Parameter	
Detector array	128-pixel
Resolution of spectra / nm	6.25
Acquisition mode	reflectance
Spectral range / nm	900-1650
Integration time / ms	10
Each spectrum scans average	60 scans
Measurement time / s	0.6

The spectral information of 57 reconstituted tobacco samples (provided by Yunan Tobacco Biological Technology Co., Ltd., Kunming, China) was obtained by using a hand-held NIR spectrometer. A polytetrafluoroethylene (PTFE) background disc was used as the spectral reference. The PTFE device was provided by Viavi Solution, Milpitas, CA, United States. Each sample was scanned for three times, and the average value of the three spectra was set as the final spectrum of the sample. The spectrum of reconstituted tobacco collected by handheld NIR spectrometer was shown in Figure 1. The 57 samples were divided in a 7:3 ratio, that is, there were 40 groups in the training set and 17 groups in the test set. The thickness of the reconstituted tobacco leaves employed in the study was about 140-300 μm. The details of NIC and VG content of the samples were shown in Table 2. The range of NIC
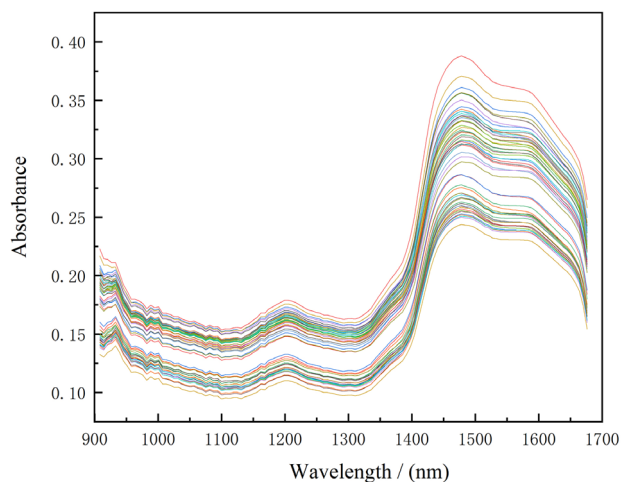

Figure 1. Near infrared spectrum of reconstituted tobacco leaves.

content was 1.48-3.06%, the average value and standard deviation were 2.15 and 0.38%, respectively. The range of VG content was 14.14-23.16%, the average value and standard deviation were 17.3 and 1.9%, respectively.

Table 2. Details of nicotine and glycerol content

Indicator	Max value / %	Min value / %	Average value / %	Standard deviation / %
NIC	3.06	1.48	2.15	0.38
VG	23.16	14.14	17.30	1.90

NIC: nicotine; VG: glycerol.

NIC and VG contents information were acquired by continuous flow and gas chromatography.[10] The continuous flow method used water to extract tobacco samples. The total phytocides (in terms of NIC) in the extract reacted with *p*-aminobenzene sulfonic acid (provided by Shanghai Aladdin Biochemical Technology Co., Ltd., Shanghai, China) and cyanide chloride (provided by Shanghai Yiji Industrial Co., Ltd., Shanghai, China), which was produced by on-line reaction of potassium cyanide and chloramine T (provided by Hebei Fangqian New Material Technology Co. Ltd., Shijiazhuang, China). The reactants were determined by colorimeter (provided by Light Analysis Technology Co., Hong Kong, China) at 460 nm. The details of the method are provided in YCT 160-2002.[11] In gas chromatography, VG was extracted from samples by methanol solution with internal label and determined by meteorological chromatograph equipped with hydrogen flame detector. The meteorological chromatograph was the GC-8890 device provided by Shandong Zhipu Information Technology Co. Ltd., Zaozhuang, China. The details of the method are provided in YCT 243-2008.[12] The rest of the devices and reagents provided by Yunnan Comtestor Co. Ltd., Kunming, China.

In this paper, the work was completed with the help of MATLAB[13] and Origin[14] software. For the SVR calculations, a MATLAB toolbox developed and described by Gunn was used.[15]

Theory of principal component regression

Principal component analysis (PCA) was mainly used to extract independent index information from the data set. There are several PCA algorithms of which non-linear iterative partial least-squares (NIPALS) and singular value decomposition (SVD) are two of the most common. PCA decomposes an **X** matrix into two smaller matrices, one of scores (**T**) and the other of loadings (**P**) as follows:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P} \qquad (1)$$

2 of 8

J. Braz. Chem. Soc. **2024**, 35, 5, e-20230174

Hence each principal component, *a*, is characterized by: (*i*) a scores vector t_a being the ath column of $\mathbf{T}$, (*ii*) a loadings vector p_a being the ath row of $\mathbf{P}$; and (*iii*) an eigenvalue g_a which may be defined by:

$$g_a = \sum_{i=1}^{I} t_{ia}^2 \tag{2}$$

Principal components (PCs) are often presented geometrically. Spectra can be represented as points in *J* dimensional space where each of the *J* axes represents the intensity at each wavelength. The first PC can be defined as the best fit straight line in this multi-dimensional space. The scores represent the distance along this line, and the loadings the direction (angle) of the straight line. If there is only one compound in a series of spectra, all the spectra will fall approximately on the straight line, since the intensity of each spectrum will relate directly to concentration. This distance is the score of the PC. If there are two components, ideally two PCs will be calculated, representing the axes of a plane.

Another important property of PCs is often loosely called orthogonality. Numerically this means that:

$$\sum_{i=1}^{I} t_{ia} t_{ib} = 0 \tag{3}$$

$$\sum_{j=1}^{I} p_{aj} p_{bj} = 0 \tag{4}$$

or $t_a \cdot t_b = 0$ and $p_a \cdot p_b = 0$ for two components *a* and *b* using vector notation. Some authors state that principal components are uncorrelated. Strictly speaking this property depends on data preprocessing, and is only true if the variables have been centred (down each column) prior to PCA. We will, however, use the terminology 'orthogonality' to refer to these properties below.

Principal components are sometimes called abstract factors, and are primarily mathematical entities. PCR uses regression (sometimes called transformation or rotation) to convert PC scores onto concentrations.[16,17]

Theory of partial least squares

Partial least squares (PLS) was a regression modeling method of multiple dependent variables to multiple independent variables. The PLS method mainly uses the following formulation:

$$\hat{\mathbf{y}}_{\mathbf{test}} = \mathbf{X}_{\mathbf{test}}^{\mathbf{T}} \mathbf{b} \tag{5}$$

where, $\mathbf{X}_{\text{test}}$ is the spectral data matrix, $\hat{\mathbf{y}}_{\text{test}}$ is the data

matrix formed by the corresponding label of $\mathbf{X}_{\text{test}}$. The decomposition renders the score matrix $\mathbf{T}$, the loading matrix $\mathbf{P}$, and the weight loading matrix $\mathbf{W}$. $\mathbf{b}$ is the vector of PLS regression coefficients obtained during the calibration step from:

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{\mathbf{T}}\mathbf{W})^{-1}\mathbf{T}^{+}\hat{\mathbf{y}}_{\mathbf{test}} \tag{6}$$

where the superscript '+' indicates the pseudoinverse operation.[18]

PLS sets the PCA method and other advantages of multiple regression analysis methods in one, because of its "response" matrix and the existence of prediction function, it can avoid some potential problems, such as the data was not in accordance with the normal distribution, data structure asymmetry. Hence, it was often used as a linear regression model in the analysis of predictive data.

Theory of support vector machine

Support vector machine (SVM) might be regarded as the perfect candidate for spectral regression purposes. A large advantage of SVM-based techniques is their ability to model nonlinear relationships. SVM has the advantage of leading to a global model that is capable of efficiently dealing with high dimensional input vectors.[19,20]

The regression of SVM can be defined by minimizing the following cost function:

$$\text{minimize: } \mathbf{L}(\mathbf{b}, e) = \frac{1}{2} \|\mathbf{b}\|^2 + c\sum_{n=1}^{N}\left(e_n + e_n^*\right), c \geq 0$$

$$\text{subject to: } y_n - \phi(\mathbf{X}_n)\mathbf{b} - b_0 \leq \varepsilon + e_n, e_n \geq 0 \tag{7}$$

$$\phi(\mathbf{X}_n)\mathbf{b} + b_0 - y_n \leq \varepsilon + e_n^*, e_n^* \geq 0$$

The cost function $\mathbf{L}$ in equation 7 consists of a 2-norm penalty on the regression vector and an error term multiplied by the error weight, c, to simultaneously minimize both the regression vector size and the prediction errors as defined in terms of some margin ε according to the two sets of inequality constraints. The problem is commonly solved by introducing Lagrange multipliers (β_n, β_n^*) for the constraints and reformulating the optimization problem in equation 7 as a quadratic programming problem. The regression vectors are then obtained from an expansion of the Lagrange multipliers multiplied by the corresponding training observations as follows:

$$\mathbf{b} = \sum_{n=1}^{N} \beta_n' \phi\left(\mathbf{X}_n\right) \tag{8}$$

J. Braz. Chem. Soc. **2024**, *35*, 5, e-20230174

3 of 8

where $\beta'_n = \beta_n - \beta_n^*$. Based on equation 8, without the bias of b_0, the prediction of inputs can be produced by the inner product as $\hat{\mathbf{y}}_{\textbf{test}} = \phi(\mathbf{X})\hat{b} = \phi(\mathbf{X})\phi(\mathbf{X})^T\beta = \mathbf{K}\beta$, where the Gram matrix $\phi(\mathbf{X})\phi(\mathbf{X})$ is defined as a dot product of inputs, and is referred to as the kernel matrix, $\mathbf{K}$, that is required to meet certain conditions (Mercer's conditions), such that the kernel matrix $\mathbf{K} = \left(K\left(\mathbf{X}_i, \mathbf{X}_j\right)\right)_{i,j=1}^{N}$ is symmetric and positive semi-definite (i.e., has non-negative eigenvalues). Among many possible options of kernel functions, the typical kernel function used in SVM is the Gaussian kernel function.[21]

Process and evaluation methods

The specific process of data processing was shown in Figure 2, which mainly consists of the following steps:

(*i*) The collected reconstituted tobacco samples were divided into two parts: training and test samples, the samples were divided randomly;

(*ii*) the handheld near-infrared spectrometer was used to collect spectroscopic data for the three samples, and the corresponding NIC and VG content information was obtained using standard chemical methods;

(*iii*) the spectroscopic data was successively preprocessed, and the characteristic spectral ranges were selected;

(*iv*) three algorithms PCR, PLS and SVM were used to build the respective training models and the optimal model was obtained by analyzing the results;

(*v*) in the subsequent procedure, the NIC and VG content of the test samples could be obtained in real time by substituting the spectral data into the training model.

Coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) were the mostly used as the evaluation indexes in the evaluation of the established spectral model. The determined coefficients were random variables sampled at random and can be used to test the reliability of the model. It was calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_m\right)^2} \tag{9}$$

where n was the sample size, y_i was the actual value of the sample, $\hat{y}_i$ was the predicted value of the sample obtained by the established model, $\bar{y}_m$ was the average value of the sample, and the R^2 value of $0 < R^2 < 1$, the R^2 closer to 1, the better the performance of the model.

The root mean square error (RMSE) can well reflect the precision of the measured object, and its calculation formula was as follows:
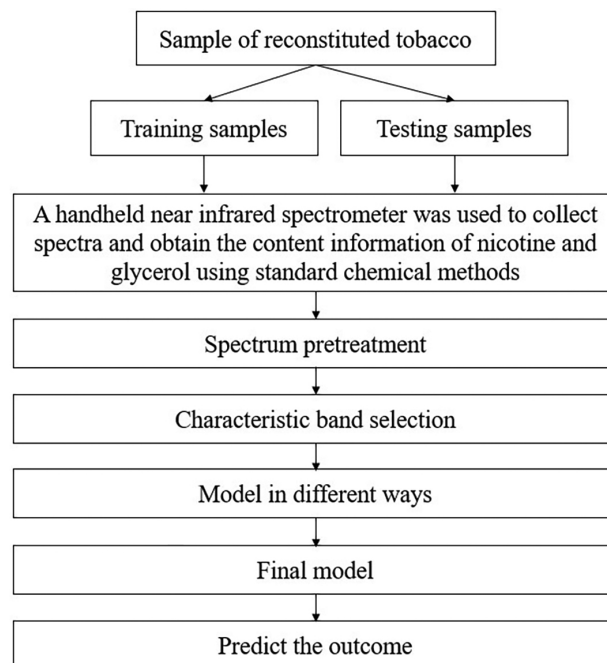

Figure 2. Procedure workflow.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{10}$$

If the prediction was more accurate, the RMSE was smaller, so the smaller the RMSE was, the better the model was. In the experiment, the root mean square error of the training set was usually denoted as RMSEC, while the root mean square error of prediction was denoted as RMSEP.

Mean absolute error (MAE) can accurately reflect the size of the actual prediction error. It was calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{11}$$

Theoretically, the smaller the MAE was, the better the model was.

In this study, R^2, RMSEC, RMSEP and MAE will be used as the evaluation indexes of the model.

Results and Discussion

Compared to desktop NIR spectroscopy instrument, the spectral data collected by a handheld NIR device have more noise and a relatively lower signal-to-noise ratio. In order to reduce the effect of noise and improve the accuracy of the prediction model, it was necessary to do some pre-processing operations on the spectra data. MSC, standard normal variate (SNV), Savitzky-

4 of 8

J. Braz. Chem. Soc. **2024**, *35*, 5, e-20230174

Golay convolution smoothing method combining with 1st derivatives (SG1) and with 2nd derivatives (SG2), were chosen in the pre-processing process. The details of four different pre-processing results were shown in Table 3. In Table 3, the optimal SVM model of NIC and VG could be obtained by using the SNV pretreatment when the coefficient of determination (R^2) was the highest and RMSEP was the lowest. The best predicted R^2 and RMSEP for NIC and VG indicators were 0.8525, 0.0484% and 0.8008, 0.0631% with the best preprocessing of SNV, respectively. It indicated that the SVM model could achieve the best performance with the SNV method. Here, a mean result was obtained by calculating the average of 20 times each result. Figure 3 was the pre-processing result and it showed that the effect of scattering of the spectrum could be eliminated effectively via SNV operation. The spectral absorption ability of information was enhanced and the signal-to-noise ratio was improved.

Table 3. Results of SVM model of nicotine and propylene glycol using different preprocessing methods

Indicator	Preprocessing method	SVM	
		R^2	RMSEP / %
NIC	NO	0.6958	0.0936
	MSC	0.8141	0.0491
	SNV	0.8525	0.0484
	SG1	0.7068	0.0743
	SG2	0.7070	0.0735
VG	NO	0.6240	0.0865
	MSC	0.7865	0.0712
	SNV	0.8008	0.0631
	SG1	0.7418	0.0692
	SG2	0.7847	0.0710

SVM: support vector machine; R^2: coefficient of determination; RMSEP: the root mean square error of prediction; NIC: nicotine; NO: no preprocessing method is used; MSC: multiplicative scatter correction; SNV: standard normal variate transform; SG1: Savitzky-Golay convolution smoothing method combining the first derivatives; SG2: Savitzky-Golay convolution smoothing method combining the second derivatives; VG: glycerol.

Table 4 showed the results of SVM model with different parameters. In Table 4, the SVM model achieved the highest R^2 and the lowest RMSEP for the NIC indicator when the parameters were set radial basis function as the kernel function type and ε-SVM (epsilon-support vector regression). The same conclusion holes for the VG indicator was achieved. The R^2 and RMSEP for NIC and VG indicator were 0.8525, 0.0484% and 0.8008, 0.0631%, respectively. Hence, the same parameters would be used for modeling in all subsequent experiments. In this study,
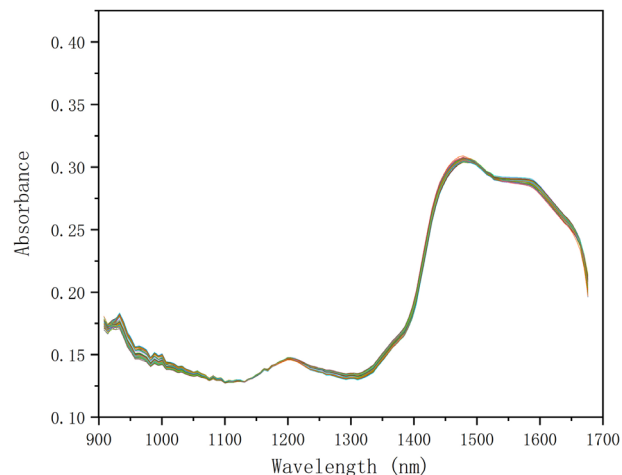

Figure 3. Spectral data after SNV preprocessing operation.

SVM was applied by choosing the ε-SVR algorithm with the following parameters: radial basis function (RBF) as kernel type, the gamma (γ) and error costs (C) for NIC and VG indicator were 0.0055, 11.3137 and 0.001, 32.

Table 5 showed R^2 and RMSEC of the training model with the optimal pretreatment and PCR, PLS, SVM algorithms. It could be seen from Table 5 that the performance of two models by using SVM algorithms was better than that of PCR and PLS algorithms. The R^2 of NIC training model by SVM algorithms was 0.9610, which was 0.237 and 0.0732 higher than that of PCR and PLS; the RMSEC of NIC training model by SVM algorithms was 0.0103%, which was 0.1914 and 0.1183% lower than that of PCR and PLS. The R^2 of VG training model by SVM algorithms was 0.9117, which was 0.1706 and 0.0373 higher than that of PCR and PLS; the RMSEC of VG training model by SVM algorithms was 0.0180%, which was 0.9467 and 0.6539% lower than that of PCR and PLS.

Table 6 showed R^2 and RMSEP of the prediction model with the optimal pretreatment and PCR, PLS, SVM algorithms. It could be seen from Table 6 that the performance of two models by using SVM algorithms was better than that of PCR and PLS algorithms. The R^2 of NIC test model by SVM algorithms was 0.8525, which was 0.232 and 0.0846 higher than that of PCR and PLS; the RMSEP of NIC test model by SVM algorithms was 0.0484%, which was 0.1897 and 0.1399% lower than that of PCR and PLS. The R^2 of VG test model by SVM algorithms was 0.8008, which was 0.142 and 0.0256 higher than that of PCR and PLS; the RMSEP of VG test model by SVM algorithms was 0.0631%, which was 1.0467 and 0.8405% lower than that of PCR and PLS. Regression by SVM can be very useful due to its ability to find nonlinear, global solutions and its ability to work with high dimensional input vectors.

J. Braz. Chem. Soc. **2024**, *35*, 5, e-20230174

5 of 8

Table 4. Results of SVM model of nicotine and propylene glycol using different parameters

Indicator	Type of kernel function	Type of SVM	R^2	RMSEP / %
NIC	linear	epsilon-SVM	0.6511	0.1117
	polynomial	epsilon-SVM	0.7222	0.0686
	radial basis function	epsilon-SVM	0.8525	0.0484
	sigmoid	epsilon-SVM	0.4925	0.1859
VG	linear	epsilon-SVM	0.6506	0.1353
	polynomial	epsilon-SVM	0.4734	0.3965
	radial basis function	epsilon-SVM	0.8008	0.0631
	sigmoid	epsilon-SVM	0.7203	0.0740

SVM: support vector machine; R^2: coefficient of determination; RMSEP: the root mean square error of prediction; NIC: nicotine; VG: glycerol.

Table 5. R^2 and RMSEC of the training model with different modeling algorithms

Indicator	Preprocessing method	PCR		PLS		SVM	
		R^2	RMSEC / %	R^2	RMSEC / %	R^2	RMSEC / %
NIC	SNV	0.7240	0.2017	0.8878	0.1286	0.9610	0.0103
VG	SNV	0.7411	0.9647	0.8744	0.6719	0.9117	0.0180

PCR: principal component regression; PLS: partial least squares; SVM: support vector machine; R^2: coefficient of determination; RMSEC: the root mean square error of the training set; NIC: nicotine; SNV: standard normal variate transform; VG: glycerol.

Table 6. R^2 and RMSEP of the prediction model with different modeling algorithms

Indicator	Preprocessing method	PCR		PLS		SVM	
		R^2	RMSEP / %	R^2	RMSEP / %	R^2	RMSEP / %
NIC	SNV	0.6205	0.2381	0.7679	0.1883	0.8525	0.0484
VG	SNV	0.6588	1.1098	0.7752	0.9036	0.8008	0.0631

PCR: principal component regression; PLS: partial least squares; SVM: support vector machine; R^2: coefficient of determination; RMSEC: the root mean square error of the training set; NIC: nicotine; SNV: standard normal variate transform; VG: glycerol.

Table 7 showed MAE of the prediction model in all cases with the optimal pretreatment. The table presented that either nicotine or glycerol leads to the same conclusion: the model built by SVM algorithm had a minimum MAE value in the same SNV preprocessing. The MAE for NIC and VG indicators were 0.0587 and 0.3474%, respectively. The MAE of NIC indicator was 0.0587, which was 0.0996 and 0.0417 lower than that of PCR and PLS; the MAE of VG indicator was 0.3474, which was 0.4588 and 0.1697 lower than that of PCR and PLS. The explanation was that, the predicted value of the model built by SVM algorithm was closer to the true value. That is, for SVM algorithm was the best to predict nictine and glycerol content than PCR and PLS.

Figure 4 showed the nicotine and glycerol estimated *vs.* nominal. Figures 4a, 4b and 4c present NIC prediction by PCR, PLS and SVM algorithms, respectively, whereas Figures 4d, 4e and 4f show VG prediction by PCR, PLS and SVM algorithms, respectively. In Figures 4a, 4b and 4c, the *x* axis represents the NIC value acquired by the

Table 7. MAE of the prediction model

Preprocessing method	Algorithm	NIC / %	VG / %
SNV	PCR	0.1583	0.8062
SNV	PLS	0.1004	0.5171
SNV	SVM	0.0587	0.3474

NIC: nicotine; VG: glycerol; SNV: standard normal variate transform; PCR: principal component regression; PLS: partial least squares; SVM: support vector machine.

continuous flow protocol and the *y* axis shows the values acquired by NIR spectroscopy. In Figures 4d, 4e and 4f, the *x* axis represents the VG value acquired by the gas chromatography protocol and the *y* axis shows the values acquired by NIR spectroscopy. These figures indicate the model was operating properly and that the spectral measurements closely fit the NIC and VG content values by the use of the SVM algorithm.

The elliptic joint confidence region (EJCR) method was applied to assess the prediction ability of the supervised
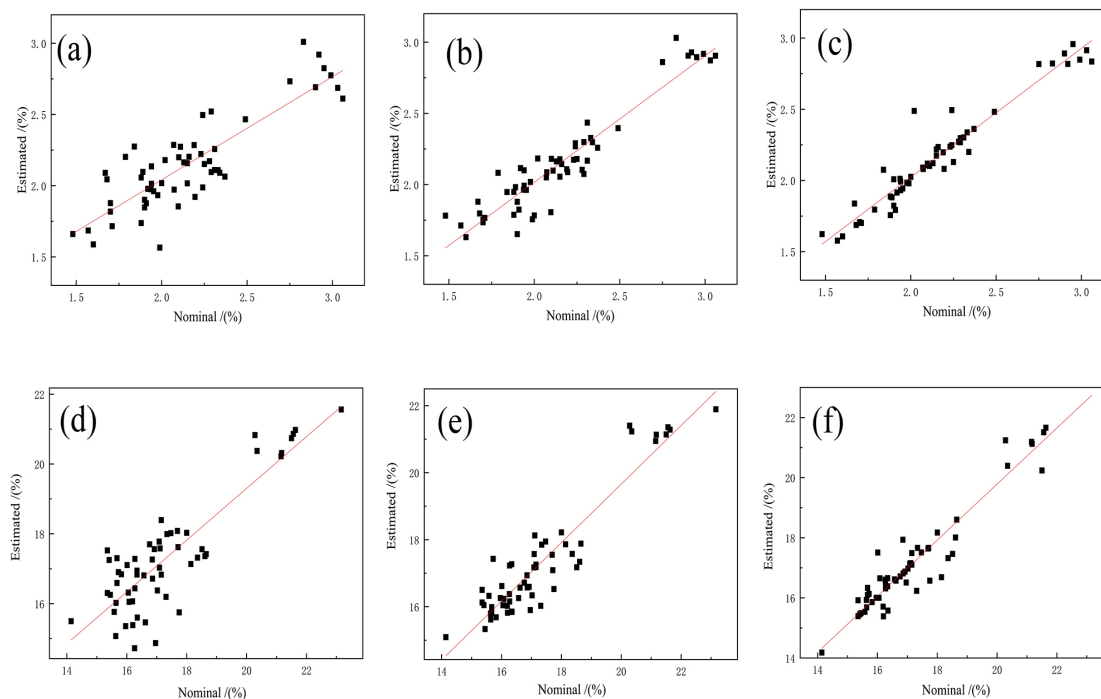
6 of 8

J. Braz. Chem. Soc. **2024**, *35*, 5, e-20230174

Figure 4. Plots of estimated *vs.* nominal values of nicotine by (a) PCR, (b) PLS, (c) SVM algorithms; glycerol by (d) PCR, (e) PLS, (f) SVM algorithms. The red lines indicate the perfect fit.

pattern recognition methods.[22] Indeed, using the EJCR, it is possible to investigate the existence of systematic errors and to evaluate the accuracy of the methodology employed.[23] Figure 5 showed the EJCR of estimated *vs.* nominal values, the 95% confidence level was chosen in each subgraph.

Figures 5a, 5b and 5c present NIC prediction by PCR, PLS and SVM algorithms, respectively, whereas Figures 5d, 5e and 5f show the VG prediction by PCR, PLS and SVM algorithms, respectively. The EJCR for all SVM models showed there were no significant differences between the
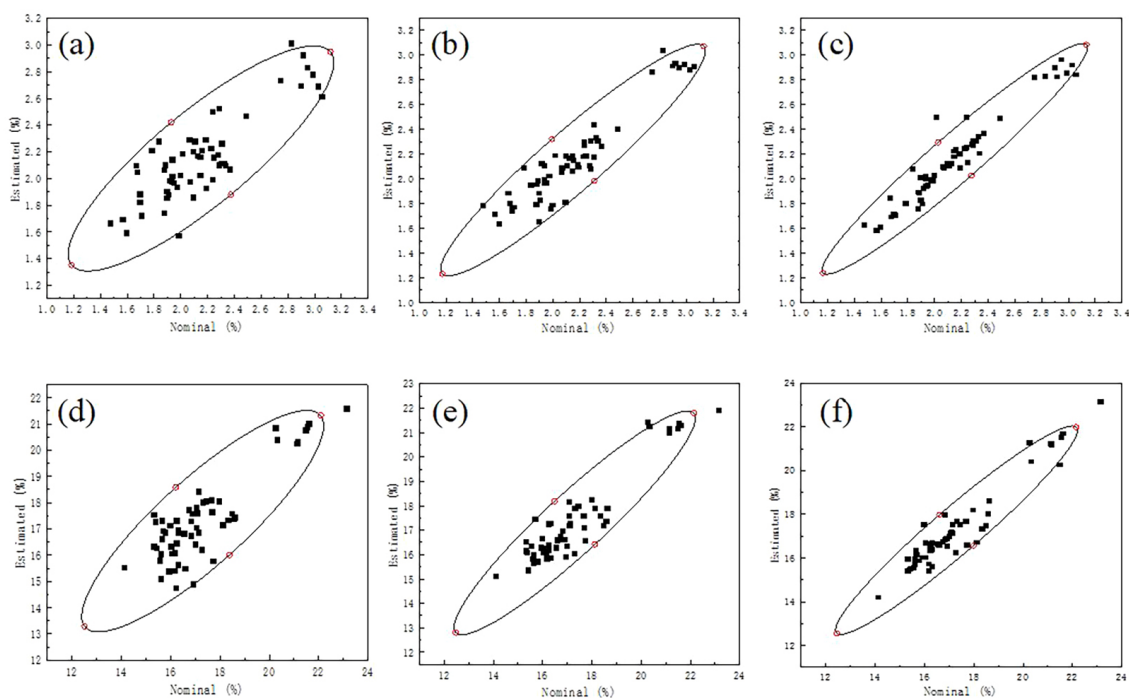

Figure 5. The elliptic joint confidence region (EJCR) of estimated *vs.* nominal values of nicotine by (a) PCR, (b) PLS, (c) SVM algorithms; glycerol by (d) PCR, (e) PLS, (f) SVM algorithms.

J. Braz. Chem. Soc. **2024**, *35*, 5, e-20230174

7 of 8

estimated values and SVM nominal values as well as that there was no evidence of bias (not shown). Therefore, all SVM models were able to quantify the NIC and VG contents in samples with excellent results.

Conclusions

Considering the advantages of a hand-held NIR spectrometer, the study focused on how to determine NIC and VG content of a reconstituted tobacco leave rapidly. A hand-held NIR spectrometer was used to collect the spectral data of reconstituted tobacco leaves and the models of NIC and VG were built by using a SVM algorithm. Besides, the performances of SVM models were also compared with PCR and PLS models. The experimental results showed that a hand-held NIR spectrometer combined with SVM algorithm can determine NIC and VG content of a reconstituted tobacco leave rapidly.

Acknowledgments

Author Contributions

Yang Qifu wrote and revised the original manuscript; Ma Kun, Zhang Xinyu, and Wu Jiaquan provided theoretical guidance and overall planning; Liang Ying and Yang Shuangyan proofread the manuscript.

References

1. Wen, Y. P.; Huang, P.; Deng, C. J.; Ying, D. F.; Gong, S.; Wang, W.; Hu, Y. N.; Zhao, G. L.; *CN pat. 112617271B* **2022**.

2. Yang, C.; Yan, G.; Qu, F.; Long, L. J.; Zhang, W. F.; Li, Q. W.; Zuo, Z. X.; Ou, M. Y.; *CN pat. 114698867A* **2022**.

3. Long, L. J.; Li, W. Q.; Qu, F.; Long, X. Z.; Gao, Z. P.; Long, J. L.; Chen, Y. M.; Chen, W. P.; Shao, F. Y.; Long, J. R.; Ou, M. Y.; Tong, F. Q.; *CN pat. 111528514B* **2022**.

4. Liu, J.; Xiang, H. Y.; Wang, B. X.; Wang, J. J.; Bai, X. L.; Lu, W.; Wu, J.; Wu, L. J.; J. *Kunming Univ. Sci. Technol., Nat. Sci. Ed.* **2015**, *40*, 108 (in Chinese). [Crossref]

5. Wang, J. J.; Lian, Y. Z.; Fan, W.; *Chinese J. Anal. Chem.* **2005**, *33*, 793 (in Chinese). [Link] accessed in October 2023

6. Li, J. S.; Li, H. J.; Lian, J. J.; Wang, Z. F.; Liu, W.; Hou, J. Q.; Zhao, J. S.; Gao, S. P.; *J. Instrum. Anal.* **2007**, *5*, 655 (in Chinese). [Crossref]

7. Li, P.; Ma, Y. J.; Ma, L.; Yang, Y. Q.; Du, G. R.; *J. Hunan Agric. Univ., Sci. Technol.* **2018**, *44*, 251 (in Chinese). [Crossref]

8. Yuan, E. W.; Yan, X. L.; Ge, J.; Zhao, D. H.; Su, L.; *Acta Agric. Jiangxi.* **2015**, *27*, 78. [Crossref]

9. Gang, W.; Shuangshuang, W.; Huaicheng, Z.; Yichen, L.; Yiming, Z.; Wei, Z.; *International Conference on Bio-Inspired Computing: Theories and Applications*, vol. 1363; Pan, L.; Pang, S.; Song, T.; Gong, F., eds.; Springer: Singapore, 2020, p. 60. [Crossref]

10. Li, M. H.; Liu, J.; Wu, J. L.; Chen, Y. C.; Zhu, T.; Qin, Q.; Yang, X.; Li, C. H.; Li, J. C.; Zhang, L. Z.; *Chem. Anal. Meterage.* **2022**, *31*, 40. [Link] accessed in October 2023

11. YC/T 160-2002: *Tobacco and Tobacco Products - Determination of Total Alkaloids - Continuous Flow Method*, State Tobacco Monopoly Administration: Beijing, 2002. [Link] accessed in October 2023

12. YC/T 243-2008: T*obacco and Tobacco Products - Determination of 1,2-Propylene Glycol and Glycerol-Gas Chromatographic Method*, State Tobacco Monopoly Administration: Beijing, 2008. [Link] accessed in October 2023

13. *Matlab*®, version R2022a; The MathWorks Inc., Natick, MA, USA, 2007.

14. *Origin*®, version 2022; OriginLab, USA, 1992.

15. Gunn, S. R.; *Support Vector Machines for Classification and Regression*; Image Speech and Intelligent Systems Group, University of Southampton, UK, 1998. [Link] accessed in October 2023.

16. Brereton, R. G.; *Analyst* **2000**, *125*, 2125. [Crossref]

17. Bro, R.; Smilde, A. K.; *Anal. Methods* **2014**, *6*, 2812. [Crossref]

18. Allegrini, F.; Olivieri, A. C.; *Anal. Chim. Acta* **2022**, *1226*, 340248. [Crossref]

19. Balabin, R. M.; Lomakina, E. I.; *Analyst* **2011**, *136*, 1703. [Crossref]

20. Thissen, U.; Pepers, M.; Üstün, B.; Melssen, W. J.; Buydens, L. M. C.; *Chemom. Intell. Lab. Syst.* **2004**, *73*, 169. [Crossref]

21. Ni, W. D.; Norgaard, L.; Mørup, M.; *Anal. Chim. Acta.* **2014**, *813*, 1. [Crossref]

22. Ding, X. X.; Ni, Y. N.; Kokot, S.; *Anal. Lett.* **2014**, *47*. [Crossref]

23. Luna, A. S.; Gonzaga, F. B.; da Rocha, W. F. C.; Lima, I. C. A.; *Spectrochim. Acta, Part B* **2018**, *139*, 20. [Crossref]