Combination of NIR spectroscopy and machine learning for monitoring chili sauce adulterated with ripened papaya

Ravipat Lapcharoensuk^{1*}, Kitticheat Danupattanan¹, Chaowarin Kanjanapornprapa¹ and Tawin Inkawee¹

¹King Mongkut's Institute of Technology Ladkrabang, Faculty of Engineering, Department of Agricultural Engineering, Post Harvest Innovation Research and Development Laboratory, Bangkok, Thailand

Abstract. This research aimed to study the combination of NIR spectroscopy and machine learning for monitoring chilli sauce adulterated with papaya smoothie. The chilli sauce was produced by the famous community enterprise of chilli sauce processing in Thailand. The ingredients of the chilli sauce consisted of 45% chilli, 25% sugar, 20% garlic, 5% vinegar, and 5% salt. The chilli sauce sample was mixed with ripened papaya (Khaek Dam variety) smoothie with 9 levels from 10 to 90 %w/w. The NIR spectra of pure chilli sauce, papaya smoothie and 9 adulterated chilli sauce samples were recorded using FT-NIR spectrometer in the wavenumber range of 12500 and 4000 cm^{-1} . Three machine learning algorithms were applied to develop a model for monitoring adulterated chilli sauce, including partial least squares regression (PLS), support vector machine (SVM), and backpropagation neural network (BPNN). All model presented performance of prediction in the validation set with $R^2_{\text{val}} = 0.99$ while RMSEP of PLS, SVM and BPNN were 1.71, 2.18 and 3.27% w/w respectively. This finding indicated that NIR spectroscopy coupled with machine learning approaches were shown to be an alternative technique to monitor papaya smoothie adulterated in chilli sauce in the global food industry.

1 Introduction

Chilli sauce is one popular flavouring which is widely consumed in Thailand as well as around the world. The taste of the chilli sauce is a combination of sour, salty, sweet and spicy. Due to the fantastic taste and satisfaction of consumers, chilli sauce is a product with high market demand. Generally, high-quality chilli sauce is produced from fresh red chilli and other ingredients such as vinegar, sugar, garlic and salt. However, for some manufacturers in Thailand, ripened papaya has always been used to adulterate chilli sauce because the physical property of ripened papaya is similar to chilli sauce, including its visual characteristic, viscosity, specific gravity and so on. Although the adulteration of papaya in chilli sauce is

* Corresponding author: ravipat.la@kmitl.ac.th

not poisonous for consumers, adulterated chilli sauce may affect the satisfaction of consumers and reduce the reliability of the product. Monitoring of adulterated chilli sauce based on a powerful method is one way to guarantee the quality of chilli sauce product.

One interesting technique for quality control and monitoring adulterated chilli sauce is near-infrared spectroscopy (NIR). NIR spectroscopy is the study of the interaction between NIR radiation (NIR: 800 – 2500 nm, i.e. 12500 – 4000 cm^{-1}) and vibration of the molecular-based on overtones and combinations, especially hydrogen bonds (C-H, O-H and N-H). NIR spectroscopy is a non-destructive, fast and environmentally friendly technique for assessing the quality of food and agricultural product [1]. However, we cannot use it directly as spectral information from a NIR spectrometer for monitoring adulterated chilli sauce because the characteristics of the spectra are very complex, broad, and overlapping. Knowledge in the field of mathematics, statistics and computer science has always been used for extracting hidden information from complex chemical data, including multivariate data analysis and chemometric technique. The typical procedures of quantitative analysis for NIR spectroscopy including pre-processing of the NIR spectra data using mathematical techniques such as smoothing, multivariate scatter correction (MSC), normalisation, derivative and others. The development of a calibration model using multivariate techniques and evaluation of performance of calibration model with statistical parameters such as coefficient of determination (R^2) and root mean square error (RMSE) is needed. The popular algorithms for regression issues include partial least squares (PLS) regression, support vector machine, artificial neural networks (ANN). These techniques are also known as machine learning in the field of computer science. The combination of NIR spectroscopy and machine learning showed successful results for monitoring food adulteration in previous research such as honey [2-5], milk [6], pepper [7], sesame oil [8], *Lonicerae Japonicae Flos* [9], soybean oil [10], *Panax notoginseng* [11] and *notoginseng* [12]. However, the NIR spectrum consists of a great number of absorbance values on all wavenumber range reaching thousands of variables. For this reason, some machine learning techniques may not be feasible in the model development process. Many types of research have suggested ways for solving this issue with the dimensionality reduction technique, such as successive projections algorithm (SPA), principle component analysis (PCA) and factor analysis (FA). Therefore, a comparison of the performance of models from creating with full spectra and the selected variable is an interesting issue for study.

All the above information brings the aim of this research, which is the development of a machine learning model with NIR spectrum data for detecting the adulteration of ripened papaya in chilli sauce. To the best of our knowledge, this is the first time that NIR spectroscopy coupled with machine learning has been applied to detect the concentration levels of ripened papaya adulterated in chilli sauce. The results will be useful to guarantee the quality of chilli sauce in both local and export markets.

2 Methodology

2.1 Samples preparation

A chilli sauce sample was purchased from the Kaset chilli sauce community enterprise, Bangphra Sub-district, Sriracha District, Chonburi Province, Thailand. The ingredients of chilli sauce consisted of 45% chilli, 25% sugar, 20% garlic, 5% vinegar, and 5% salt. A ripened papaya (Khaek Dam variety) was obtained from a local market. A papaya sample was peeled with stainless steel knives and its flesh was blended using a blending machine for 5 min. The adulterated chilli sauce samples were prepared by mixing papaya smoothie in chilli sauce at 9 levels from 10 and 90% w/w. A sample at each level of adulteration was

contained in a beaker (250 ml) and stirred with a stirring rod for 30 min. A beaker of each sample was closed with a lid and plastic wrap until the session of image and NIR spectra collection.

2.2 NIR spectra collection

Before the NIR spectra scanning, each sample was poured from a beaker into 10 vials (20 mm of diameter and 43 mm of height). Therefore, the total number of NIR spectra was 110 spectra. NIR spectra of samples were measured using an FT-NIR spectrometer (MPA, Bruker Ltd., Germany) in a wavenumber range of 12500-4000 cm^{-1} (800-2500 nm) with a resolution of 8 cm^{-1} . The spectrum of each sample was recorded in an interactance mode and reported with absorbance value ($\log 1/R$). Each spectrum had a product of 32 internal scanning.

2.3 Overall precision test

The precision of the experiment was evaluated and reported in terms of repeatability and reproducibility. Repeatability is a precision test of an instrument under the same measurement condition. In this research, the same sample of pure chilli sauce was scanned NIR spectra with 10 repetitions under the controlled condition (i.e. 25°C) for reporting repeatability of NIR spectrometer. The repeatability was reported with average and standard deviation (SD). In the other hand, reproducibility explained all the procedures of the experiment can be reproduced in its entirety. In this research, the variance of the experiment might be obtained from the non-homogeneity from the mixing process of adulterant samples. Therefore, reproducibility was analysed with the scanning of NIR spectra on the 10 different parts of adulterant chilli sauce at all levels of adulteration. The reproducibility was described by the average SD value for each level of adulteration. The absorbance peak at 6900 cm^{-1} (absorbance peaks of water (Osborne and Fearn, 1986)) was selected to report repeatability and reproducibility because the NIR absorbance of water always obtains variance, even in a controllable experiment.

2.4 Machine learning approaches

Regression models were developed from machine learning approaches, including partial least squares regression (PLS), support vector machine (SVM), and backpropagation neural network (BPNN). Due to the number of NIR absorbance values on all wavenumber range reaching thousands of variables, it might not be feasible for the calculation of some machine learning techniques such as BPNN. Therefore, NIR spectra data should be pre-processed with dimensionality reduction technique. In this research, principal component analysis (PCA) was used for feature extraction of NIR spectra data before training BPNN. However, the BPNN were trained from full spectra and the principal components (PCs) from PCA for comparing performance of both procedures. In process of PCA, NIR spectra data was transformed from the original spectra into a set of PCs. After that, the PCs with two largest proportion of variance were employed for training BPNN. This proposal presented successful results in many previous pieces of research [14-17]. On the other hand, PLS and SVM algorithms were trained by using full NIR spectra. The sample set was split into calibration (80%) and validation set (20%) with stratified sampling method. A calibration set was used to establish the machine learning models, which optimum parameter of each algorithm was searched with the GridSearchCV command of the Scikit-learn module (Version 0.22) [18]. The potential of the model was reported on the coefficient of determination of calibration and validation set (R_{cal}^2 and R_{val}^2 respectively), root mean square error of calibration (RMSEC), root mean square error of prediction (RMSEP). Machine learning procedure was

implemented in Python (Version 3.7.4) [19] using the Scikit-learn module (Version 0.22) [18].

3 Results and discussions

3.1 NIR spectra of samples

Figure 1 shows the NIR spectra of pure chili sauce, pure papaya smoothie and adulterated chili sauce (20% w/w (C80 P20), 40% w/w (C60 P40), 60% w/w (C60 P40) and 80% w/w (C20 P80)). The explicit peaks were achieved at 10310, 8695, 6900 and 5100 cm^{-1} i.e., 970, 1150, 1450 and 1960 nm respectively). The vibration band of water (H_2O) appeared as two peaks on NIR spectra at 6900 and 10310 cm^{-1} which were the first and second overtone of O-H stretching [20]. The peak at 8695 cm^{-1} related with the second overtone of C-H stretching of $\text{HC}=\text{CH}$ [20]. The remaining peak at 5100 cm^{-1} corresponded to the combination of NIR vibration band of CONH_2 [20]. NIR absorbance of samples decreased significantly according to the levels of adulteration of papaya smoothie in chili sauce. This indicates the possibility of monitoring chilli sauce adulterated with papaya smoothie with NIR spectroscopy coupled with machine learning.

Fig. 1. NIR spectra of pure chili sauce, pure papaya smoothie and adulterated chili sauce with papaya smoothie at 20% w/w (C80 P20), 40% w/w (C60 P40), 60% w/w (C40 P60) and 80% w/w (C20 P80).

3.2 Overall precision test

The repeatability and reproducibility of this experiment were 0.002 and 0.060% w/w, respectively. Statistically, the SD value explains the amount of variation in the data set. A low value of the standard deviation indicates that the data points tend to be close to the expected value. In this research, repeatability and reproducibility showed low standard deviation, indicating that the NIR spectrometer and mixing process were precise under both similar and different conditions.

3.3 Principle component analysis

Figure 2 shows a scatter plot of PC-1 and PC-2 factors from the PCA technique. The percentage of the variance for PC-1 was 89.5% and PC-2 was 10%. Chilli sauce and adulterated chilli sauce at 10 to 40 %w/w of papaya smoothie (C90 P10, C80 P20, C70 P30, C60 P40) distributed on the positive axis of PC-1, while the remaining samples (papaya

smoothie, adulterated chilli sauce at 50 to 90% w/w) were scattered on the negative axis of PC-1.

The PC loading plot from the PCA algorithm is shown in Figure 3. Important information of the NIR vibration band can be explained by the PC loading plot from PCA analysis. Permanent peaks of PC loading were obtained at 10310, 8695, 7170, 6900, 6025 and 5100 cm^{-1} (970, 1150, 1395, 1450, 1660 and 1960 nm). The NIR vibration band of H_2O molecule appeared on the peaks at 10130 and 6900 cm^{-1} which are the first overtone and second overtone stretching of O-H [20]. Two peaks around 8695 and 6025 cm^{-1} were assigned to first overtone and second overtone of C-H stretching in capsaicin molecule [20-21]. The peak at 5100 cm^{-1} also related vibration band of capsaicin molecule which is the combination of N-H stretching [20-21].

Fig. 2. Scatter plot of PC-1 and PC-2 factors from the PCA algorithm of pure chilli sauce, pure papaya smoothie and adulterated chilli sauce with papaya smoothie at 10, 20, 30, 40, 50, 60, 70, 80 and 90% w/w (C90 P10, C80 P20, C70 P30, C60 P40, C50 P50, C40 P60, C30 P70, C20 P80, C10 P80 respectively).

Fig. 3. PC-loading from NIR spectra of samples.

3.4 Performance of machine learning approaches

Table 1 shows the results for monitoring of chilli sauce adulterated with papaya smoothie with NIR spectroscopy coupled with machine learning algorithms. All techniques except development of BNPP with full spectra showed high performance of the test set with $R^2_{val} = 0.99$ while RMSEP of these algorithms were 1.71, 2.18 and 3.27 for PLS, SVM and BPNN, respectively. However, the BNPP with full spectra presented $R^2_{val} = 0.81$ and RMSEP = 12.01 which represented that this model is not suitable to apply for evaluating adulteration of

papaya in chili sauce. This point indicated that NIR spectra data should be reduced in dimensionality before training neural network algorithm. The scatter plot of prediction using the 4 machine learning approaches is shown in Figure 4.

Table 1. the results of monitoring of chili sauce adulterated with papaya smoothie with NIR spectroscopy coupled with machine learning.

Machine learning algorithm	R^2_{cal}	RMSEC (% w/w)	R^2_{val}	RMSEP (% w/w)
PLS (number of factors = 2)	0.99	1.03	0.99	1.71
SVM	0.99	0.49	0.99	2.18
BPNN (full spectrum)	0.81	14.45	0.80	12.01
BPNN (PCs)	0.99	3.32	0.99	3.27

Fig. 4. Scatter plots of prediction of the level of adulterated papaya smoothie in chilli sauce using machine learning approaches: (a) calibration set from PLS; (b) validation set from PLS; (c) calibration set from SVM; (d) validation set from SVM; (e) calibration set from BPNN with full spectra; (f) validation set from BPNN with full spectra; (g) calibration set from BPNN with PCs; (h) validation set from BPNN with PCs.

Figure 5 shows the regression coefficient plot of the PLS model, which was developed with the full range of NIR spectra. Permanent peaks obtained were at 12250, 11450, 11110, 10660, 10310, 9430, 8680, 8370, 7350, 6370, 5865, 5100 and 4710 cm^{-1} (815, 874, 900, 938, 970, 1060, 1150, 1195, 1360, 1570, 1960 and 2132 nm). Peaks of regression coefficient represent the effect of the molecular vibration band on the PLS model at these wavenumbers⁶. For this research, these evident peaks related to the molecular vibrations of the consequential chemical composition of chilli and papaya including water, starch, and capsaicin. The regression coefficient peak at 10310 cm^{-1} is the second overtone of O-H stretching of water. The obvious peaks at 11110, 8680, 8370, 7352 and 5865 cm^{-1} represented the first overtone, second overtone and combination of C-H stretching of methyl ($-\text{CH}_3$) [6, 20] in the capsaicin molecule. In addition, NIR vibration band of capsaicin molecule evinced obviously at peaks of 12250, 9430, 6370, 5100 and 4710 cm^{-1} which assigned to the first overtone, second overtone, third overtone and combination vibration band of N-H and C-O. The remaining peak at around 10660 cm^{-1} related to the first overtone and combination of C-H stretching of methylene ($-\text{CH}_2$) [6, 20].

Fig. 5. Regression coefficient plot form PLS algorithm.

4 Conclusions

The combination of near-infrared (NIR) spectroscopy coupled with machine learning was proposed to monitoring adulteration of ripened papaya in chilli sauce. The PLS model developed from raw spectra showed the highest potential for detecting adulterated chilli sauce in a validation set with $R_{\text{val}}^2 = 0.99$ and $\text{RMSEP} = 1.71\%$ w/w. In the case of developing with PCs from PCA analysis, BPNN presented R_{val}^2 and RMSEP at 0.99 and 3.27, respectively. The prediction results of the BPNN models creating with PCs was better than the BPNN models creating with full spectra. This point indicated guideline for training neural network algorithm with NIR spectra data. These results could conclude that NIR spectroscopy coupled with machine learning can be used as an alternative technique to guarantee the quality of chilli sauce in the global industry.

References

1. R. Lapcharoensuk, N. Nakawajana, J. Innov. Opt. Health Sci. **11(1)**, 1750019 (2018)
2. M. J. A. González, M. F. González, E. E. Bellido, M. Palma, G. F. Barbero, Talanta **203**, 235-241 (2019)
3. M. F. González, E. E. Bellido, L. G. Cueto, M. Palma, G. F. Barbero, Talanta **188**, 288-292 (2018)
4. L. Chen, X. Xue, Z. Ye, J. Zhou, J. Zhao, Food Chem. **128(4)**, 1110-1114 (2011)
5. X. Zhu, S. Li, Y. Shan, Z. Zhang, F. Liu, J. Food Eng. **101(1)**, 92-97 (2010)
6. R. Lapcharoensuk, J. Chaianate, S. Winichai, A. Phetnak, Proc. SPIE **11205**, 1120516 (2019)
7. R. Lapcharoensuk, S. Chalachai, S. Sinjaru, P. Singsri, J. Hongwiangjan, N. Yaemphochai, IOP Conf. Ser. Earth Environ. Sci. **301**, 012068 (2019)
8. H. Chen, Z. Lin, C. Tan, Vib. Spectrosc. **99**, 178-183 (2018)
9. J. Zhao, P. Cui, H. Liu, C. Wang, T. Suo, Infrared Phys. Technol. **104**, 103139 (2020)
10. C. G. Pereira, A. I. N. Leite, J. Andrade, M. J. V. Bell, V. Anjos, LWT-Food Sci. Technol. **107**, 1-8 (2019)
11. P. Liu, J. Wang, Q. Li, J. Gao, X. Bian, Spectrochim. Acta. A Mol. Biomol. Spectrosc. **206**, 23-30 (2019)
12. H. Chen, C. Tan, Z. Lin, H. Li, Spectrochim. Acta. A Mol. Biomol. Spectrosc. **211**, 280-286 (2019)
13. W. S. Rasband, ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, <http://rsb.info.nih.gov/ij/> (1997-2019)
14. K. Igor, R. Glen, H. Charles, J. Near Infrared Spectrosc. **15**, 21-28 (2007)
15. C. Borggaard, H. H. Thodberg, Anal. Chem. **64**, 545-551 (1992)
16. T. Naes, K. Kvaal, T. Isaksson, C. Miller, **1**, 1-11 (1993)
17. J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, P. S. Meltzer, Nat. Med. **7**, 673-679 (2001)
18. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, J. Mach. Learn. Res. **12**, 2825-2830 (2011)
19. Python Software Foundation, <https://www.python.org/> (2001-2019)
20. B. G. Osborne, T. Fearn, *Theory of near infrared spectrophotometry* (In: Near Infrared Spectroscopy in Food Analysis, London, Longman Science & Technical, 1993)
21. M. Iwamoto, R. K. Cho, U. Jun, I. Kyuei, *Japan Soc. Food Sci. Technol.* **31(2)**, 120-125 (1984)