Hyperspectral detection of walnut protein contents based on improved whale optimized algorithm

Yao Zhang[1], Zezhong Tian[1], Wenqiang Ma[1], Man Zhang[1], Liling Yang[2*]

(1. *Key Laboratory of Smart Agriculture System Integration, Ministry of Education, China Agricultural University, Beijing 100083, China*; 2. *Agricultural Mechanization Institute, Xinjiang Academy of Agricultural Sciences, Urumqi 830091, China*)

Abstract: Nondestructive and accurate estimation of walnut kernel protein content is important for food quality grading and profitability improvement of walnut packinghouses. Hyperspectral image technology provides potential solutions for walnuts nutrients detection by obtaining both spectral and textural information. However, the redundancy and large computation of spectral data prevent the widespread application of hyperspectral technology for high throughput evaluation. For walnut kernel protein inversion from hyperspectral image, this study proposed a novel feature selection method, which is named as improved whale optimized algorithm (IWOA). In the IWOA, a comprehensive feature selection criterion was applied in the iterative process, which fully considered the relevance of spectra information with target variables, representative ability of the selected wavebands to entire spectra, and redundancy of the selected wavebands. Especially in the relevance with target variables, the amplitude and shape characteristics of the spectra were both taken into consideration. Eight wavelengths around 996, 1225, 1232, 1377, 1552, 1600, 1691 and 1700 nm were then selected as the sensitive wavelengths to walnut protein. These wavelengths showed good correlation with certain chemical compounds related to protein contents mechanistically. Then three protein prediction models were established. After analysis and comparison, the model based on the selected wavelengths got better results with the one based on the full spectrum. Compared to the models based on solely spectral information, the model that combine spectral and textural information outperformed and got the best prediction results. The R^2 in the calibration group was 0.9047, and the root mean square errors (*RMSE*) was 11.1382 g/kg. In the validation group, the R^2 was 0.8537, and the *RMSE* was 18.9288 g/kg. The results demonstrated that the combination of the selected wavelengths through the IWOA with the textural characteristics could effectively estimate walnut protein contents. And the proposed method can be extended to the detection and inversion of other nutritional variables of nuts.

Keywords: walnut protein, hyperspectral image, whale optimized algorithm, feature selection, textural indicator

DOI: 10.25165/j.ijabe.20221506.7179

Citation: Zhang Y, Tian Z Z, Ma W Q, Zhang M, Yang L L. Hyperspectral detection of walnut protein contents based on improved whale optimized algorithm. Int J Agric & Biol Eng, 2022; 15(6): 235–241.

1 Introduction

Walnut is an important woody oil crop and is highly valued for its nutritional walnut kernels. Walnut kernels contain 15%-22% of protein, of which more than 96% is human absorbable protein, which is the highest compared with soybeans, peanuts, almonds, hazelnuts and eggs[1]. Protein content of walnut kernels is one of the most important factors determining perceived quality and ultimate price of walnut products. Nondestructive and accurate obtaining the walnut kernel protein content is crucially important for walnut kernel quality grading, and would assure the competitiveness and profitability of the walnut industry.

The traditional methods of kernel protein detection are chemical measurements, which are destructive and have the potential for environmental contamination. Considering the demands in practice, it is more necessary to develop a fast and efficient method to accomplish the walnut kernel protein content detection. NIR spectroscopy, on the basis of electromagnetic characteristics of matter, makes it possible to detect protein content of food quickly and noninvasive, which could ensure the safety of food production during whole inspection processes[2,3]. However, spectroscopic measurements are generally obtained from a limited area (i.e., point measurement using ASD FieldSpec FR spectroradiometer (Analytical Spectral Devices, Inc., Boulder, CO, USA))[4,5] or from prepared samples of specific size (i.e. limited area measurement using MATRIX-I type of FT-NIR analyzer with a rotating sample pool (Bruker Optical Company, Germany))[6]. They do not provide image information of objects such as texture or location information, which is important in many food inspection applications[7,8].

The advent of hyperspectral technology has enabled the perfect integration of spectroscopy as well as imaging analysis techniques, providing more information about the targeted objects, and provides a possibility to further improve the accuracy of composition content detection of targets. Hyperspectral data containing full-bands radiation information could describe various characteristics associated with the biochemical and physiological traits of targets[9-13]. But the existing data redundancy and band autocorrelation in hyperspectral information could always lead to an increase of computation complexity and the incorrect test results.

Received date: 2021-11-09 **Accepted date:** 2022-05-20

Biographies: Yao Zhang, Associate Professor, research interest: agricultural information engineering, Email: zhangyao@cau.edu.cn; **Zezhong Tian,** Master, research interest: agricultural remote sensing engineering, Email: tzz@cau.edu.cn; **Wenqiang Ma,** Associate Professor, research interest: agricultural intelligent equipment, Email: mwq4530@163.com; **Man Zhang,** Professor, research interest: agricultural information engineering, Email: cauzm@cau.edu.cn.

***Corresponding author: Liling Yang**, Professor, research interest: agricultural intelligent equipment. Xinjiang Academy of Agricultural Sciences, No. 403, Nanchang Road, Urumqi 830091, China. Tel: +86-13999821292, Email: 411450712@qq.com.

In recent years, the deep learning, as a state-of-the-art technique, has been extensively applied in the field of hyperspectral image processing, in which the features could be learned automatically according to the targeted tasks[14-18]. And large number of labeled samples is desirable to ensure the stability of the deep learning models. However, in actual productions, collecting such large amounts of labeled data is expensive or generally impossible[19]. Furthermore, the limited amounts of labelled sample set constrained the wide application and performance of deep learning algorithm in feature extraction and composition content determination[20,21]. Therefore, it is important to fully exploit the effective features contained in the hyperspectral images and reduce the dependence of feature selection on the volume of the dataset.

Therefore, the advantages of features selection method which is embedded in existing experience become obvious, because it no longer needs additional large amounts of labelled samples for training. Among the feature selection approaches, swarm-based optimization algorithms mimicking biological or physical phenomena could be used to solve complex feature selections[22-25]. Typical swarm-based optimization algorithms, such as genetic algorithm and particle swarm optimization algorithm can improve their stochastic capability by setting some parameters. But there is no guarantee that these algorithms will be able to search globally and jump out of the local optimum in hyperspectral feature selection. Moreover, some parameters such as mutation rate and particle velocity can seriously affect the quality of the solution, which need to be set with a lot of known experience[26]. A novel nature-inspired meta-heuristic optimization algorithm, called Whale Optimization Algorithm (WOA), which mimics the social behavior of humpback whales and inspired by the bubble-net hunting strategy was introduced in 2016[27]. WOA, including three operators to simulate the search for prey, encircling prey, and bubble-net foraging behavior of humpback whales, shows advantages over other state-of-the-art meta-heuristic methods in exploration, exploitation, local optima avoidance, and convergence behavior[25,28,29].

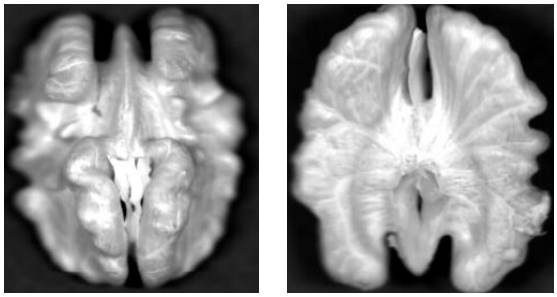
While in the process of feature selection using WOA, most studies only adopted random principle to set up whale foraging behaviors, which makes the selected features mostly dependent on the performance of WOA and does not fully consider the contribution of hyperspectral characteristics including amplitude and shape information in feature selection[30]. Besides, for variable selection, there are three factors which are quite important to be considered. One is discrepancies with the selected wavebands; the other one is the representative ability of the whole spectral information and the last one is the correlation level with the target variable. To address the above issues, this study presented an improved WOA algorithm combined with hyperspectral characteristics and feature selection criteria.

This study takes walnut as the research object and aims to propose a novel inversion method for walnut protein contents estimation based on hyperspectral information. The paper is organized as follows. Section 2 introduces the walnut samples and hyper-spectrum determination in this study. Detailed description about the proposed optimized WOA feature selection method and random forest regression model are described in Section 3. Results of selected protein sensitive wavebands, validation of protein inversions based on different training dataset are presented in Section 4. The advantages and limitations of the proposed method are described in Section 5. The paper concludes in Section 6 with a summary of the results.

2 Materials

2.1 Walnut samples and pre-treatment

In this study, Xinjiang 'Wen185' walnuts were selected as target objects. The experimental samples, with water content of 7%, was stored at 4℃ for about 5 months before the experiment. After manually shell breaking, 30 walnut kernel samples were collected. The front and back side of the half kernel samples are shown in Figure 1. After the hyperspectral image is collected by hyperspectral imaging equipment, the protein content was measured by Kjeltec automatic nitrogen analyzer (Foss, Denmark). Each sample was crushed and selected 0.3 g for the analyzer. After nitrification, the samples were titrated by distillation to obtain the total nitrogen content. Then the total nitrogen content was multiplied by a conversion factor (6.25) to calculate the protein content of the walnuts in accordance with the rule of the National Food Safety Standard GB5009.5-2010.



a. Front side b. Back side

Figure 1 Front and back side of the half kernel samples

2.2 Hyper-spectrum determination

Hyperspectral image of walnut kernel was measured in a laboratory by hyperspectral imager (Gaia sorter, Zhuoli Hanguang Company, Beijing, China), which is mainly composed of imaging spectrometer (V10E), lens (OL23), CCD (LT365), uniform light source (2 sets of tungsten bromide lamps), electric control mobile platform, computer and software control system. Warm up first after startup to eliminate the impact caused by baseline drift. After preheating, focusing the lens and adjusting the moving speed of the platform to avoid image distortion. The image acquisition software Spectrum View was used to collect the imaging information of walnut kernel. The hyperspectral measurement ranges were 863-1704 nm and 382-1027 nm respectively. The spectral resolutions in the two spectral ranges were 3.2 and 0.84 nm. In order to eliminate the noise influence caused by uneven illumination, measurement environment and dark current of the instrument, it is necessary to collect white background information (I_w) and black background information (I_b) by using standard whiteboard and lens cover respectively before collecting the hyperspectral image of the sample. Then according to Equation (1), the corrected hyperspectral image (I) was obtained from original hyperspectral image (I_o).

$$I = \frac{I_o - I_b}{I_w - I_b} \quad (1)$$

where, i is the corrected image; I_w is the while background image; I_b is the black background image; I_o is the original image.

3 Methodology

3.1 Improved whale optimization algorithm (WOA)

Hyperspectral feature selection using original WOA[27] mostly relies on the performance of WOA in feature selection and does not make full use of hyperspectral characteristics to achieve more in-depth valid information mining. Therefore, in the process of

WOA optimization, a comprehensive evaluation criterion for hyperspectral characteristic variable selection is proposed, which takes into accounts three aspects: the degree of correlation between the wavelength information and the target (protein content), the redundancy between the selected wavelengths, and the ability of the selected wavelengths to represent the full spectrum. The wavelength information includes the amplitude and shape information that is unique to hyperspectral information. Based on the above principles, Pearson Correlation Coefficient is used to calculate the above three indicators. The general expression is shown below, which was used as the fitness function of the algorithm:

$$EC = Norm_0^1(Co(x_i, y)) + Norm_0^1\left(\sum_{j=1}^{N} Rep(x_{ai}, x_{aj})\right) - Norm_0^1\left(\sum_{j=1}^{SN} Red(x_{ai}, x_{aj})\right) \quad (2)$$

$$Co(x_i, y) = Norm_0^1\left(Co\_amp(x_{ai}, y)\right) + Norm_0^1\left(Co\_shp(s_{si}, y)\right) \quad (3)$$

where, EC represents the results of evaluation criterion; C_o represents the correlation degree; R_{ep} indicates the representative degree; R_{ed} indicates the redundancy of the selected wavelengths; C_{o_amp} is the correlation degree between spectral amplitude information and target; C_{o_shp} is the correlation degree between spectral shape information and target; x is the spectral information, x_a is the spectral amplitude; x_s is the spectral shape value; y is the protein contents; N is the number of all wavelengths; SN is the number of selected wavelengths.

In the above expression, the spectral shape information is the angle between the extension direction of the spectral curve within adjacent bands and the horizontal direction. First, the hyperspectral reflectance was normalized from 1 to the number of wavelengths. The angle was then computed as shown in Equation (11). Subsequently, all the angle information of the entire curve was obtained and used to describe the shape characteristics of the hyperspectral curve.

$$x_{si} = \arctan\left(\frac{xa'_{i+1} - xa'_i}{\lambda_{i+1} - \lambda_i}\right) \quad (4)$$

where, xa'_{i+1} and xa'_i are the normalized reflectance at wavelength i+1 and i, respectively; λ_{i+1} and λ_i are the wavelength of i+1 and i, respectively.

The feature selection evaluation criteria are effectively confused with the original WOA in terms of whale hunting behavior and fitness function in the optimization algorithm, which could ensure that the whales could jump out of the local optimum in the search process and comprehensively evaluate the selected features in terms of representativeness, relevance and redundancy. In addition, the unique amplitude and shape information in the hyperspectral information is effectively used to fully mine the hyperspectral information.

The process of the improved WOA algorithm is as follows:

Step1 Initialize the whale population location, set the whale population size and the maximum number of iterations

Step2 Calculate the fitness (as Equation (2)) of all whales in the group, and record the one with the highest fitness value as the location of the prey X_p

Step3 Update parameters a, A, C, l and p, their setting reference citation[27]

Step4 Update the location of each whale

if p<0.5

if $|A|$<1, encirclement contraction predation method

update the position of each whale according to Equation(5)

$$X(t+1) = EC(t) \times X_{rand}(t) - A \cdot D \quad (5)$$

else if |A|≥1, random exploration of predation

update the position of each whale according to Equation (6)

$$X(t+1) = EC(t) \times X_{best}(t) - A \cdot D \quad (6)$$

end

else if p≥0.5, perform bubble-net feeding method

update the position of each whale according to Equation (7)

$$X(t+1) = D' \cdot e^{bl} \cdot \cos(2\pi l) + EC(t) \times X_p(t) \quad (7)$$

end

Step5 Calculate the fitness of all whales in the population as Equation (2), and record the one with the largest fitness value as the optimal solution. If the iteration stopping condition is met or the maximum number of iterations is reached, the algorithm stops; otherwise, it returns to Step3.

3.2 Gray level co-occurrence Matrix (GLCM)

The recurrence of pixel grayscale in spatial location forms the texture of the image, and GLCM is a description of the joint distribution of two pixels grayscale with spatial location relationships. Haralick et al.[31] proposed GLCM to characterize texture features. In this research, the GLCM of the walnut images were calculated first, then four texture statistical indicators were calculated. Contrast (*Con*) was used to describe the sharpness of the textures and its calculation as shown in Equation (8). *Dis* and *Homo* can reflect the dissimilarity and homogeneity of the textures and the local textural variation separately, they are calculated by Equations (9) and (10). Energy (*E*), an indicator used to describe the uniformity of the distribution of greyscale and the coarseness of the texture, is calculated by Equation (11).

$$Con = \sum_{i=1}^{k}\sum_{j=1}^{k}(\mathrm{GLCM})_{i,j}^2 (i-j)^2 \quad (8)$$

$$Dis = \sum_{i=1}^{k}\sum_{j=1}^{k}(\mathrm{GLCM})_{i,j} |i-j| \quad (9)$$

$$Homo = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{(\mathrm{GLCM})_{i,j}^2}{1+(i-j)^2} \quad (10)$$

$$E = \sum_{i=1}^{k}\sum_{j=1}^{k}\sqrt{(\mathrm{GLCM})_{i,j}^2} \quad (11)$$

3.3 Random forest

The random forest algorithm consists of multiple decision trees, and it is an efficient and reliable integrated learning method with good tolerance to sample data outliers and noise. The randomness of random forest is reflected in the sample randomness and feature randomness. The random forest algorithm adopts the Bagging strategy to train decision trees, which means that m subsamples are randomly selected from the original training set for building m decision trees, and then the random forest will randomly select some variables from all independent variables as the nodes of the decision trees to reduce the correlation between each decision tree and make the decision process more multivariate.

Because the samples that are not selected each time under the Bagging strategy form the out-of-bag data set, there is no need to set aside extra data for cross-validation, and the random sample selection also reduces the computational effort, while the final decision result collects the information of all decision trees to ensure the prediction accuracy of the model.

4 Results

4.1 Walnut NIR spectra and selected wavelengths

Walnut kernels are mainly composed of fat, protein, water, sugar and other trace elements. The electromagnetic wave signal in the near-infrared domain contained the most characteristic bands related to the above components. Therefore, due to the high sensitivity of the spectral signal in the NIR region, only hyperspectral images in the range of 863-1704 nm were selected for subsequent analysis in this study. Figure 2 shows the NIR spectral characteristic extracted from the hyperspectral images of 30 walnut kernel samples. Figure 2 reveals that the variation tendency of the walnut spectral curves of the different samples becomes similar with various protein contents. The spectral information at both ends of the curve (before 870 nm and after 1680 nm) contained a considerable amount of noise because of the vibration of the measurement system. Two obvious peaks around 1210 nm and 1470 nm could be observed, which are caused by water content. It could be concluded, except for the reflectance peaks of water, the reflectance peaks of other components were not obvious, and further processing of the spectra was required.

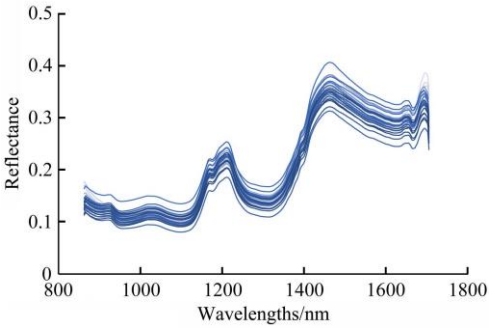


Figure 2 Spectral curves of walnuts samples

As described in Section 3.1, the proposed improved WOA for feature wavelengths selection of walnut kernel is calculated and visualized through self-developed software. The positions of selected wavebands in the whole spectral range using improved WOA algorithm are shown in Figure 3. The eight wavelengths around 996, 1225, 1232, 1377, 1552, 1600, 1691 and 1700 nm were selected. The previous publications revealed that the 996 nm was corresponding to protein contents in the identification of rough rice species and years by visible/near-infrared[32]. Tallada et al.[33] found the usual significant absorption peaks between 1175-1225 for protein through mean spectral profile of the 87 maize seed samples. Nagao et al.[34] determine the fat content in meats using a combination of absorbances at 1208 nm and 1230 nm. The optical transmittance spectrum recorded for the grown glycine phosphite single crystal shows an absorption edge at 1377 nm in the upper wavelength region. Hence it is clear that the grown glycine phosphite crystal has a transmittance window at 1377 nm with nearly about 100% transparency[35]. Yadav et al.[36] utilize this tapered fiber optic biosensor, operating at 1550 nm, for the detection of protein concentration. Capus and Cockcroft[37] measured the refractive indices of protein solution with different concentrations using an Abbe type refractometer at a wavelength of 1700 nm. It could be concluded that, after using the improved WOA wavebands selection method, the selected wavelengths were scattered in the whole spectral range. It means that this method reduced the autocorrelation and redundancy of the selected wavelengths. And these selected wavelengths could represent the whole spectral information to a certain extent. Besides, throughout the previous studies, most characteristic bands selected through the improved WOA selection methods in this research have correlation relationship with certain chemical compounds related to protein contents in walnut kernels mechanistically.

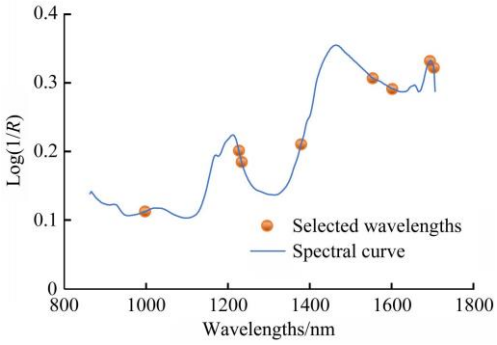


Figure 3 Spectral curve and selected wavelengths of walnuts protein

4.2 Protein estimation using selected wavelengths

To verify the protein predictive ability of the selected sensitive wavelengths, this study firstly extracted spectral curves of the interested region of front side and back side of the walnut kernels and calculated the averages as the targeted full-spectrum information. The eight wavelengths namely, 996, 1225, 1232, 1377, 1552, 1600, 1691 and 1700 nm were then selected. Two datasets including full-spectral information in the range of 862-1710 nm and the eight selected wavelengths mentioned above were created for the following modelling and comparison. The protein contents of 30 soil samples in the dataset were in the range of 12.5%-20%.

The RF models based on full-spectral information and the eight selected wavelengths for walnut kernel protein content estimation were established separately. Samples were divided into two groups, that is, 20 samples were under the calibration group and the remaining 10 samples were under the validation group.

According to the established models, the 1:1 relationship diagram were drawn between the prediction and observation to demonstrate the reliability and consistency of the selected models. The calibration and validation results of the RF model based on full-spectrum was shown in Figure 4. And the walnut kernel protein prediction results based on the eight wavelengths was shown in Figure 5.

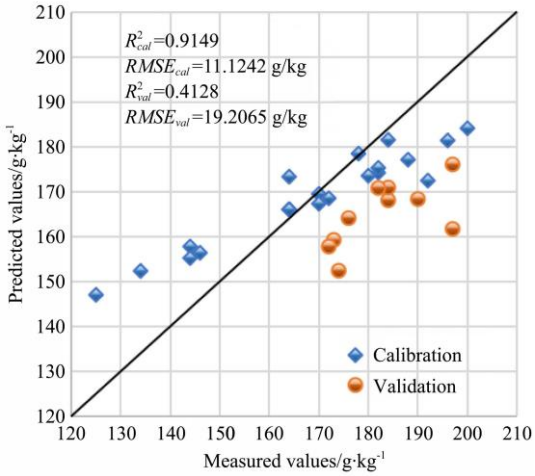


Figure 4 Calibration and validation of walnut kernel protein prediction of the RF regression models based on full spectral information

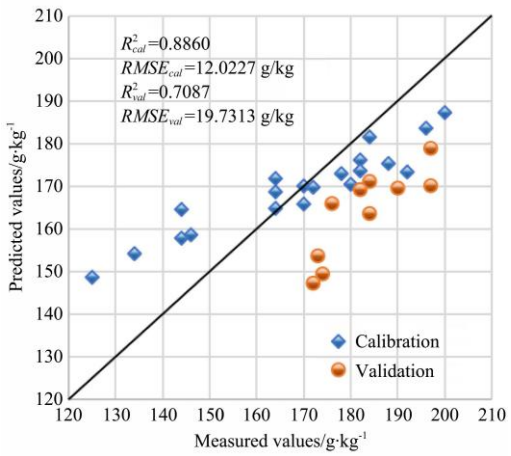


Figure 5 Calibration and validation of walnut kernel protein prediction of the RF regression models based on eight selected wavelengths

The calibration R^2 of the models based on the full-spectrum and eight selected wavelengths reached 0.9149 and 0.8860, and the root mean square errors of calibration ($RMSE_{cal}$) were 11.1242 g/kg and 12.0227 g/kg. The validation R^2 reached 0.4128 and 0.7087, and the root mean square error of prediction ($RMSE_{val}$) were 19.2065 g/kg and 19.7313 g/kg. It could be concluded that the walnut kernel protein prediction based on the eight selected wavelengths obtained better inversion results with the one based on the full spectrum, the comparative conclusion will be verified under other regression models in Section 5.

4.3 Protein estimation using the combination of spectral and texture information

In order to further increase the walnut kernel protein prediction accuracy, the texture characteristics of the front side and back side of the kernels were considered in the following modelling. As illustrated in Section 3.2, four texture indicators including contrast, dissimilarity, homogeneity and energy were used to construct a mixed dataset with the eight selected wavelengths to retrieve the protein contents of walnut kernels. 20 samples were used to calibrate the model and the remaining 10 samples were the validation group. In the RF regression model, the calibration and validation results of walnut kernel protein inversion based on the mixed dataset containing the eight wavelengths and four texture indicators is shown in Figure 6.

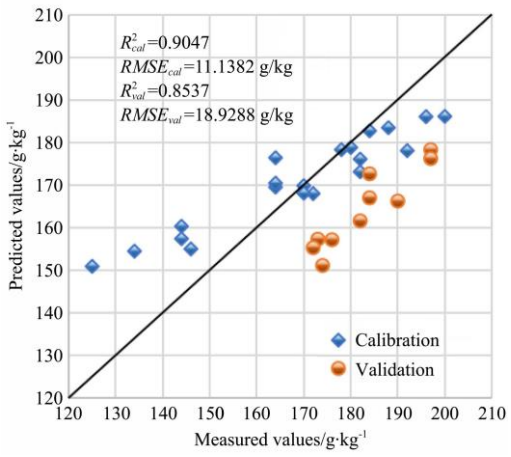


Figure 6 Calibration and validation of walnut kernel protein prediction of the RF regression models based on the mixed dataset

Compared with the model established by solely spectral information, the addition of texture information did not significantly improve the calibration modeling accuracy, but further improved the validation accuracy of protein content of walnuts. The validation R^2 of the models based on the mixed dataset increased to 0.8537. And the $RMSE_{val}$ was 18.9288 g/kg. However, comprehensively observed from Figure 4 to Figure 6, in validation process, the predicted values are lower than the measured values. Although the mixed dataset improved this situation, it still existed.

5 Discussion

Prior works have made various attempts to inverse nut nutritional factors using hyperspectral technology[14,16,38,39]. However, the universal characteristic wavelength selection method for protein content inversion is strongly needed. In this study, an improved WOA method that perfectly combines swarm intelligence and feature selection criteria, was proposed for identifying the sensitive wavelengths from the aspects of mechanism and predictive ability of walnut kernel protein. The advantages and limitations of this proposed method would be discussed as follows.

(1) The proposed wavebands selection method is on the basis of WOA, which has the advantages in exploration, exploitation, local optima avoidance, and convergence behavior[28,29]. To further improve the performance of WOA in feature selection, a comprehensive criterion was used to merge with the WOA to fully utilize spectral information and advantage of WOA in feature selection. In this criterion, three aspects, including the degree of correlation between the wavelength information and the target (protein content), the redundancy between the selected wavelengths, and the ability of the selected wavelengths to represent the full spectrum were taken into consideration in the whole process of feature selection. Besides, it should be noted that, the shape information, which is unique to hyperspectral information is also considered in this study to fully mine the NIR spectral characteristics. Most selected wavelengths using this proposed method were correlated with protein contents in walnut kernels mechanistically and supported by the previous studies[32,34-37]. This proposed method could be also applied to other areas. And further validations on other nutritional factors of nuts need to be conducted in the future. The results of this study could be transferred to the nuts packinghouses and processing industries.

(2) To further verify the validity of the selected features as well as the RF model in the process of walnut kernel protein contents inversions, two other commonly used Machine Learning (ML) regression models, including support vector machine (SVM) and back propagation neural network (BPNN) were selected as the comparison models. Subsequently, the three different models based on the full spectrum, eight selected wavelengths and a mixed dataset containing selected spectral information and texture features were established separately. The inversion accuracies of the three models are listed in Table 1. The results demonstrate that for three different regression models, the models based on the eight selected wavelengths have better or comparable performance with the one based on full spectrum. And the involvement of texture feature metrics further increased the protein inversion accuracy. The literature shows that the external epidermis of walnut kernels is developed from the growth of seed coat, which affected the asparagine synthesis[40]. Therefore, the textural characteristics of walnut kernels can reflect the internal quality of walnut kernels to some extent. Besides, among these three models, the RF model has the best inversion performance with the

same input. In particular, the RF model based on spectral and texture mixture features got best results, the R^2 is 0.8537 and RMSE is 18.9288 g/kg.

Table 1 Protein inversion results for different datasets under three ML models

Models	Datasets	R^2	*RMSE*/g kg^{-1}
SVM	Full spectra	0.2077	16.5557
	Selected wavelengths	0.2144	16.4917
	Mixed spectral and textural dataset	0.5517	16.4360
BPNN	Full spectra	0.4409	42.6338
	Selected wavelengths	0.4543	38.1880
	Mixed spectral and textural dataset	0.6795	29.6923
RF	Full spectra	0.4128	19.2065
	Selected wavelengths	0.7087	19.7313
	Mixed spectral and textural dataset	0.8537	18.9288

(3) The research proposed an innovative spectral feature extraction method and provided convincing evidence that walnut kernel protein contents could be estimated through the extracted features. However, some limitations are worth noting. The small amount of sample size limited the application of the existed algorithms including deep learning algorithms in this study, which constrained the undertaking of more comparative analyses. As for the protein inversion results in this study, from Figures 4-6, it could be observed that the drawback of the established models is that the predicted protein content values are lower than the actual values in the validation process, which may be caused by the limited training size. In future work, to further validate the generalization of the selected features and model, more samples covering different sizes, different species, different colors, should be included. After the expansion of the dataset, more comparative analyses with existing methods should be conducted to verify the superiority of the proposed method in this study. Besides, future work should also be addressed to evaluate whether the proposed hybrid method is suitable for other nut species in protein inversion.

6 Conclusions

To effectively utilize the hyperspectral information of walnut kernel samples, a novel method merged the WOA and feature selection criteria was innovatively proposed to screen the sensitive wavebands of walnut protein. The obtained sensitive wavebands were then mixed with the texture indicators to predict walnut protein contents. The main conclusions are as follows:

(1) After wavelength selection using the proposed improved WOA method, eight wavelengths, including 996, 1225, 1232, 1377, 1552, 1600, 1691, and 1700 nm were determined as protein content sensitive wavebands. According to the previous literatures, all eight wavelengths had correlation relationship with certain chemical compounds related to protein contents in walnut kernels mechanistically, which verified the effectiveness of the improved WOA in wavelength selection.

(2) The accuracies of the RF regression model based on the selected wavebands achieved better precision with the full spectral regression models. In addition, the model based on the combination of selected wavelengths and texture indicators reached a highest accuracy in walnut protein contents prediction. All the results of the models indicated the effectiveness of the sensitive wavebands selected using improved WOA method in this research. And the full use of hyperspectral information performed well with high predictive ability in predicting the protein content.

Acknowledgements

This study was supported by the Science and Technology Innovation Key Cultivation Project of Xinjiang Academy of Agricultural Sciences (Grant No. xjkcpy-004).

[References]

[1] Mao X Y, Hua Y F. Composition, structure and functional properties of protein concentrates and isolates produced from walnut (*Juglans regia* L.). International Journal of Molecular Sciences, 2012; 13(2): 1561–1581.

[2] Mohammadi M T, Razavi S M A, Taghizadeh M. Applications of hyperspectral imaging in grains and nuts quality and safety assessment: a review. Journal of Food Measurement and Characterization, 2013; 7(3): 129–140.

[3] Vohland M, Besold J, Hill J, Fründ H. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma, 2011; 166(1): 198–205.

[4] Gholizadeh A, Neumann C, Chabrillat S, Wesemael B, Castaldi F, Borůvka L. Soil organic carbon estimation using VNIR–SWIR spectroscopy: The effect of multiple sensors and scanning conditions. Soil and Tillage Research, 2021; 211: 105017. doi: 10.1016/j.still.2021.105017.

[5] Pullanagari R R, Dehghan-Shoar M, Yule I J, Bhatia N. Field spectroscopy of canopy nitrogen concentration in temperate grasslands using a convolutional neural network. Remote Sensing of Environment, 2021; 257: 112353. doi: 10.1016/j.rse.2021.112353.

[6] An X F, Li M Z, Zheng L H, Liu Y M, Sun H. A portable soil nitrogen detector based on NIRS. Precision Agriculture, 2014; 15(1): 3–16.

[7] Ramirez-Paredes J P, Hernandez-Belmonte U H. Visual quality assessment of malting barley using color, shape and texture descriptors. Computers and Electronics in Agriculture, 2020; 168: 105110. doi:10.1016/j.compag.2019.105110.

[8] Zheng C X, Sun D W, Zheng L Y. Recent applications of image texture for evaluation of food qualities—A review. Trends in Food Science & Technology, 2006; 17(3): 113–128.

[9] Aviara N A, Liberty J T, Olatunbosun O S, Shoyombo H A, Oyeniyi S K. Potential application of hyperspectral imaging in food grain quality inspection, evaluation and control during bulk storage. Journal of Agriculture and Food Research, 2022; 8: 100288. doi: 10.1016/j.jafr.2022.100288.

[10] Caporaso N, Whitworth M B, Fisk I D. Total lipid prediction in single intact cocoa beans by hyperspectral chemical imaging. Food Chemistry, 2021; 344: 128663. doi: 10.1016/j.foodchem.2020.128663.

[11] Feng L, Wu B H, Zhu S S, He Y, Zhang C. Application of visible/infrared spectroscopy and hyperspectral imaging with machine learning techniques for identifying food varieties and geographical origins. Frontiers in Nutrition, 2021; 8: 680357. doi: 10.3389/fnut.2021.680357.

[12] Khamsopha D, Woranitta S, Teerachaichayut S. Utilizing near infrared hyperspectral imaging for quantitatively predicting adulteration in tapioca starch. Food Control, 2021; 123: 107781. doi: 10.1016/j.foodcont.2020.107781.

[13] Zhang C, Liu F, Kong W W, He Y. Application of visible and near-infrared hyperspectral imaging to determine soluble protein content in oilseed rape leaves. Sensors, 2015; 15(7): 16576–16588.

[14] Engstrom O C G, Dreier E S, Pedersen K S. Predicting protein content in grain using hyperspectral deep learning. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, BC, Canada: IEEE, 2021; pp.1372–1380.

[15] Gomes V, Mendes-Ferreira A, Melo-Pinto P. Application of hyperspectral imaging and deep learning for robust prediction of sugar and pH levels in wine grape berries. Sensors, 2021; 21(10): 3459.

[16] Liu Y S, Zhou S B, Han W, Li C, Liu W X, Qiu Z F, et al. Detection of adulteration in infant formula based on ensemble convolutional neural network and near-infrared spectroscopy. Foods, 2021; 10(4): 785. doi: 10.3390/FOODS10040785.

[17] Wang L, Liu H G, Li T, Li J Q, Wang Y Z. Verified the rapid evaluation of the edible safety of wild porcini mushrooms, using deep learning and PLS‐DA. Journal of the Science of Food and Agriculture, 2022; 102(4): 1531–1539.

[18] Zhang X L, Yang J, Lin T, Ying Y B. Food and agro-product quality evaluation based on spectroscopy and deep learning: A review. Trends in Food Science & Technology, 2021; 112: 431–441.

[19] Zeng J, Guo Y, Han Y Q, Li Z M, Yang Z X, Chai Q Q, et al. A review of

the discriminant analysis methods for food quality based on near-infrared spectroscopy and pattern recognition. Molecules, 2021; 26(3): 749. doi: 10.3390/molecules26030749.

[20] Kamilaris A, Prenafeta-Boldú F X. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 2018; 147: 70–90.

[21] Wang P, Fan E, Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognition Letters, 2021; 141: 61–67.

[22] Brezočnik L, Fister I, Podgorelec V. Swarm intelligence algorithms for feature selection: A review. Applied Sciences, 2018; 8(9): 1521. doi: 10.3390/app8091521.

[23] Kumar A, Jaiswal A. Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on twitter. Multimedia Tools and Applications, 2019; 78(20): 29529–29553.

[24] Nguyen B H, Xue B, Zhang M J. A survey on swarm intelligence approaches to feature selection in data mining. Swarm and Evolutionary Computation, 2020, 54: 100663. doi: 10.1016/j.swevo.2020.100663.

[25] Zhang Y, Lee W S, Li M Z, et al. Non-destructive recognition and classification of citrus fruit blemishes based on ant colony optimized spectral information. Postharvest Biology and Technology, 2018; 143: 119–128.

[26] Trivedi I N, Bhoye M, Bhesdadiya R H, Jangir P, Jangir N, Kumar A. An emission constraint environment dispatch problem solution with microgrid using whale optimization algorithm. 2016 National Power Systems Conference (NPSC). Bhubaneswar, India: IEEE, 2016; pp.1–6.

[27] Mirjalili S, Lewis A. The whale optimization algorithm. Advances in Engineering Software, 2016; 95: 51–67.

[28] Mafarja M M, Mirjalili S. Hybrid whale optimization algorithm with simulated annealing for feature selection. Neurocomputing, 2017; 260: 302–312.

[29] Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Applied Soft Computing, 2018; 62: 441–453.

[30] Wang M W, Jia Z T, Luo J W, Chen M L, Wang S P, Ye Z W. A hyperspectral image classification method based on weight wavelet kernel joint sparse representation ensemble and β-whale optimization algorithm. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021; 14: 2535–2550.

[31] Haralick R M, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, 1973; SMC-3(6): 610–621.

[32] Shao Y N, He Y, Cao F. Identification of rough rice species and years by visible/near-infrared reflectance spectroscopy. 2006 International Conference on Computational Intelligence and Security. Guangzhou, China: IEEE, 2006; pp.988–991.

[33] Tallada J G, Palacios-Rojas N, Armstrong P R. Prediction of maize seed attributes using a rapid single kernel near infrared instrument. Journal of Cereal Science, 2009; 50(3): 381–387.

[34] Nagao A, Uozumi J, Iwamoto M, et al. Determination of fat content in meats by near-infrared reflectance spectroscopy. Journal of Japan Oil Chemists' Society, 1985; 34(4): 257–261.

[35] Devi K R, Srinivasan K. Synthesis, growth, morphology and characterization of ferroelectric glycine phosphite single crystals. Crystal Research and Technology, 2011; 46(12): 1265–1272.

[36] Yadav T K, Narayanaswamy R, Abu Bakar M H, Mustapha Kamil Y, Mahdi M A. Single mode tapered fiber-optic interferometer based refractive index sensor and its application to protein sensing. Optics Express, 2014; 22(19): 22802. doi: 10.1364/OE.22.022802.

[37] Capus J M, Cockcroft M G. A new technique for investigating surface flow in metal-working processes. Nature, 1954; 173(4409): 821–821.

[38] Nogales-Bueno J, Baca-Bocanegra B, Hernández-Hierro J M, Garcia R, Barroso J M, Heredia F J, et al. Assessment of total fat and fatty acids in walnuts using near-infrared hyperspectral imaging. Frontiers in Plant Science, 2021; 12: 729880. doi: 10.3389/fpls.2021.729880.

[39] Zhao X, Wang W, Ni X Z, Chu X, Li Y F, Sun C P. Evaluation of Near-Infrared Hyperspectral Imaging for Detection of Peanut and Walnut Powders in Whole Wheat Flour. Applied Sciences, 2018; 8(7): 1076. doi: 10.3390/app8071076.

[40] Rao G D, Sui J K, Zhang J G. Metabolomics reveals significant variations in metabolites and correlations regarding the maturation of walnuts (*Juglans regia* L.). Biology Open, 2016; 5(6): 829–836.