Original Article
Food Science and Technology

ISSN 0101-2061 (Print)
ISSN 1678-457X (Online)

DOI: https://doi.org/10.1590/fst.46522

Rapid determination of water content in potato tubers based on hyperspectral images and machine learning algorithms

Zhiyong ZOU[1], Qingsong WU[1], Jie CHEN[1], Tao LONG[1], Jian WANG[1], Man ZHOU[2], Yongpeng ZHAO[1], Tingjiang YU[3], Yinfan WANG[4], Lijia XU[1*]

Abstract

This study investigated the hyperspectral reflectance response of time series generated during oven drying to changes in the moisture content of potato tubers. Seventeen preprocessing methods were used to eliminate the influence of spectral noise on the spectral characteristic curve. Algorithms such as CatBoost, LightGBM, and XGBoost are used to obtain the first 40 effective characteristic spectra of hyperspectral images, which reduces the redundancy of data and improves the prediction accuracy. The water content prediction model of potato tubers was established by using the selected characteristic bands. The results showed that the combined model based on Lasso and XGBoost algorithm had the strongest prediction ability. The best model is MF-Lasso-XGBoost, which has R^2 value of 0.8908, Rmse of 0.0610, Mdae of 0.0389, and R^2_{cv} of 0.8448. This research can provide reference for the detection of potato moisture content and theoretical basis for the development of crop moisture detector.

Keywords: feature extraction; hyperspectral image; machine learning; moisture content; potato tuber.

Practical Application: Prediction of water content in potato tubers by hyperspectral technology.

1 Introduction

The potato (Solanum tuberosum) is the fourth largest food crop globally after wheat, rice, and maize (Habig et al., 2018). Potatoes have high nutritional contents, such as various vitamins and minerals, including vitamin C, vitamin B6, niacin, folic acid, potassium, iron, and magnesium, as well as starch and water (Zhu et al., 2022). The processing of raw potato tubers to obtain various target products is important due to their nutritional and agronomic value (Nikzad et al., 2021; Pereira et al., 2021). Potatoes are used widely to produce flakes, flour, and other potato-based foodstuffs. Dehydration treatment is required to improve the stability of the potatoes and food products depend greatly on the moisture content of the raw potatoes used for to reduce the microbial activity during processing procedures. Thus, a rapid and effective method is required to effectively monitor the dehydration process and control the moisture content (Calderón et al., 2021). Currently, the main focus of the food drying industry is the qualitative or quantitative relationship between moisture content and product quality, rather than intelligent dynamic control of the production process. Therefore, accurate estimation of moisture content is essential to establish a reliable relationship between quality attributes and products. New moisture content analysis methods will be very important in the potato processing industry.

In recent years, machine learning methods have been found to be effective in predicting the content of food ingredients using spectral data (Hou et al., 2022). Machine learning is an extension of mathematical statistics and computer science and includes many statistical models and computer program algorithms. Liu et al. (2018) used hyperspectral technology to study the water content of potato leaves at different leaf positions, and predicted their water content through machine learning algorithms As machine learning algorithms mature, new versions of machine learning algorithms have emerged for predicting structure, folding, binding, and even catalytic activity, the main purpose of which is to process the accumulated information about mutants and their functional properties (Wang et al., 2021). Zhao et al. (2021) used a machine learning algorithm to predict the chlorophyll content of potato crops based on visible light and near-infrared spectroscopy. Zhang et al. (2022) performed a comprehensive analysis of photosynthetic pigments and SPADs by combining spectral and multispectral imaging techniques with different machine learning algorithms. Zheng et al. (2018) established a model for estimating chlorophyll content in potato leaves at the red edge position, with an R^2 of 0.87. Hou et al. (2022) used Fourier transform infrared spectroscopy and machine learning to predict the amino acid content of insects, and the analysis of insect spectral data through machine learning proved to be able to predict amino acid content. Therefore, this study provides guidance for nondestructive testing of potato water content.

Hyperspectral imaging combines the spectrum and image of the target object at the same time to accurately capture the spectral data and image information of each pixel in the image

Received 20 Apr., 2022
Accepted 04 June, 2022
[1] *College of Mechanical and Electrical Engineering, Sichuan Agricultural University, Ya'an, P. R. China*
[2] *College of Food Sciences, Sichuan Agricultural University, Ya'an, P. R. China*
[3] *State Energy Dadu River Waterfall Ditch Hydroelectric Power Plant, Ya'an, China*
[4] *College of Electrical and Electronic Engineering, Harbin University of Science and Technology, Har'bin, China*
*Corresponding author: xulijia@sicau.edu.cn

(Liu et al., 2022). In recent years, hyperspectral imaging and visualization techniques have been applied in agriculture for drought monitoring and the control of diseases and insect pests (Sun et al., 2019). Gerhards et al. (2016) successfully applied the hyperspectral reflectance data obtained from the potato crop canopy to predict the moisture contents of potato plants with high accuracy. However, the single algorithm employed in that study might not be optimal for predicting the moisture contents of potato tubers. Moisture-sensitive spectral and vegetation indexes based on spectral absorption and transmission techniques have also been applied to estimate the moisture contents of plants. For instance, the spectra acquired at 800, 1323, and 1423 nm were identified as moisture-sensitive wavelengths and a multiple linear regression (LR) model was established for predicting the moisture content of corn leaves by using the difference vegetation index (1423 nm and 800 nm) and transmission spectra at 1323 nm and 1058 nm (Sun et al., 2018). Das et al. (2020) extracted spectral indices at 1391 and 1830 nm as moisture-sensitive wavelengths and developed ratio vegetation indexes (RVI, R1391, R1830) and normalized difference spectral indices (NDSI, R139, R1830) to simultaneously measure the relative water and microelement contents of rice.

The overall goal of this study was to determine the feasibility of using hyperspectral imaging to monitor dynamic changes in potato tuber water content. Compared with the previous single method, 17 different spectral data analysis methods were used to optimize the process and improve the accuracy of the results. Reliable typical water-sensitive spectral wavelengths related to tuber moisture content were obtained by machine learning, improving the accuracy of models used to predict potato tuber moisture content. It provides new methods and ideas for the prediction of potato moisture content.

2 Material and methods

2.1 Materials

In total, this study used 200 potatoes sampled from one potato variety (*Cv.* Hezuo-88) in Yunnan province, China. The fresh potato tubers were carefully rinsed with water. Next, 104 potato samples with length > 3 cm, width > 2 cm, height > 2 cm, no damage to the skin, no deformities, and no signs of germination were selected to measure the moisture contents. A square blade measuring 15 mm × 15 mm (Jiechenuo Tech. China) was used to cut a single potato into rectangular parallelepiped tuber samples (width × height = 15 mm × 30 mm). All of the samples were stored in a refrigerator at 4 °C for 24 h to prevent moisture losses and browning after cutting. A temperature of 4 °C is the best for storage because it effectively slows down the activities of enzymes in agricultural food products (Su & Sun, 2016).

2.2 Experimental design and implementation

After 24 h, the tubers were labeled individually. The weight of each sample was recorded and controlled within the range from 10.50-11.60 g by blade micro-cutting (Santos et al., 2020). Hyperspectral data and the weight of each sample were collected in a time series during the heating process. The first hyperspectral image was labeled as t_0 before heating and oven drying. In total, 105 samples were treated together by drying in an oven (Midea, PT2531) at 120 °C. Hyperspectral photographing and weight measurements were performed at 45 min, 105 min, and 165 min, where the hyperspectral data were labeled as $t_{45\ min}$, $t_{105\ min}$, and $t_{165\ min}$, respectively, and the samples weights as M0, M45, M105, and M165. The change in water content is shown in Figure 1. All of the samples were then oven dried at 105 °C for 24 h until a constant weight (M_d). The formula for calculating the tuber moisture content is as Equation 1 (Su & Sun, 2016):

$$M_c = \frac{M_t - M_d}{M_t} \times 100\% \quad (1)$$

The moisture content (M_c) was calculated based on the difference between each weight recorded and M_d, and by dividing by M_t.

2.3 Hyperspectral imaging system and image correction

An Image-λ series hyperspectral camera (Zolli Hanguang Co. Ltd, Beijing) with a spectrum acquisition range of 387–1035 nm, resolution of 2.8 nm, 256 bands, and 1344 × 1024 pixels was used

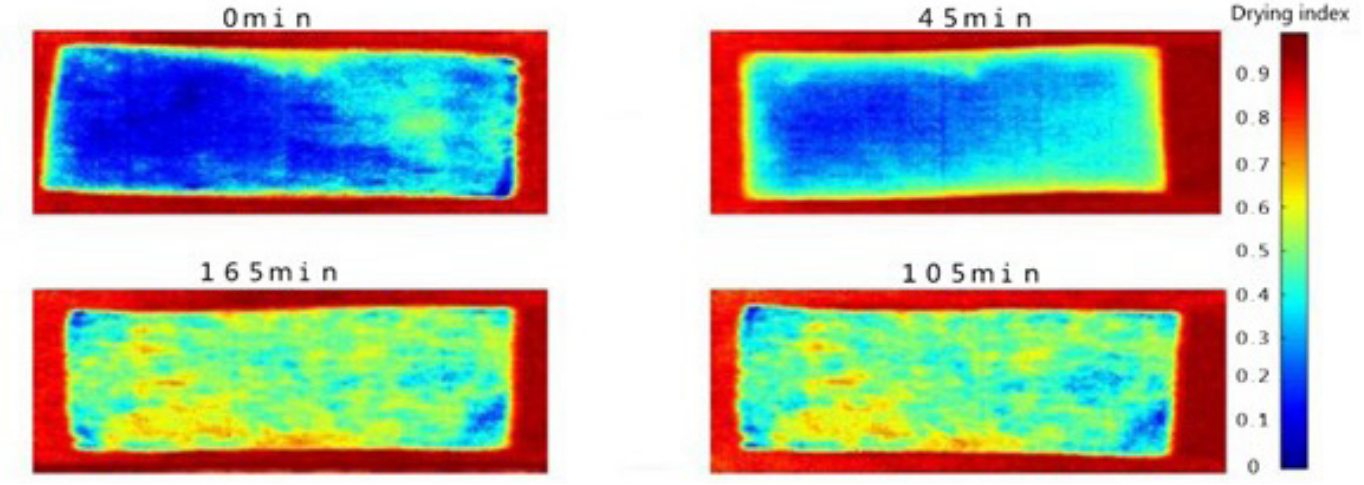

Figure 1. Visualization of potato cut water content.

to establish the laboratory platform. The distance between the hyperspectral camera and sample was set to 17 cm. The moving speed of the mobile loading platform was set to 0.5 cm s^{-1} and the measurement speed was < 60 s per sample. The exposure time for the hyperspectral camera was set to 5 ms. The scanning start position was set to 120 mm and the actual length of the scanning line was 100 mm. A block diagram illustrating the hyperspectral imaging acquisition system used in this study is shown in Figure 2.

This experiment uses ENVI5.1 (Exelis Visual Iformation Slution Is) to select ROI. Use the software to open the raw format file, use the oval tool to circle the potatoes, extract the data covering the entire potato sample, calculate the image ROI of each sample and calculate the average value of the spectrum of all pixels in the region as the sample information for the final spectral value. Among them, spectral data were extracted from potato samples of 0 min, 45 min, 105 min, 165 min and pure dry matter, respectively, with a total of 520 spectral data.

Hyperspectral imaging systems respond to various light source intensities at different wavelengths and the impact of noise is more severe when the intensity of the light source is weaker. The dark current in the camera produces significant noise, which cannot be avoided in hyperspectral images. The acquired hyperspectral images were corrected in black and white. After capturing the last image and without changing any parameters, the lens of the hyperspectral camera was aligned with a standard whiteboard to obtain a standard white frame image W. A black frame image B was obtained by covering the camera lens with a lid (Shao et al., 2022). The corrected image R was then obtained using Equation 2.

$$R = \frac{I - B}{W - B} \tag{2}$$

In Equation 2, I denote the original hyperspectral image.

The corrected hyperspectral images were then subjected to subsequent data analysis with ENVI4.8 software (Exelis Visual Information Solutions, USA). A region of interest (70 × 70 pixels) in each image was used to calculate the average spectral reflectance values from the 256 bands, which were then employed as the original hyperspectral data for predicting the moisture contents of the potato tubers.

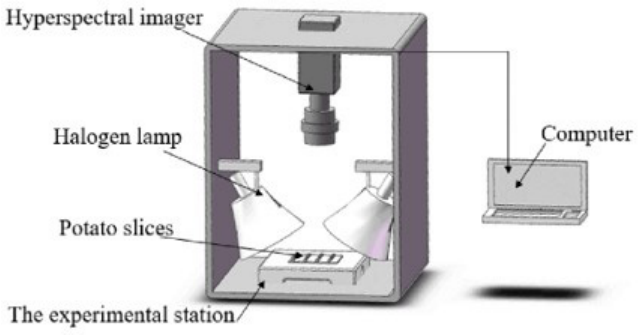

Figure 2. Block diagram illustrating the hyperspectral image acquisition system employed to study potato tubers.

2.4 Hyperspectral data preprocessing

In order to eliminate the effects of factors that had no relationships with the moisture content in the hyperspectral spectrum information, 17 data preprocessing methods were employed to eliminate the noise present in the original spectral data and to identify outliers in the box plots of the moisture contents. such as first derivative (FD), second derivative (SD), box smoothing (BS), L2 norm normalizationL2 (L2NN), moving average method (MAM), multiplicative scatter correction (MSC), min-max standardization (MMS), anti-cotangent normalization (CAN), wavelet threshold denoising (WTD), logarithmic transformation normalization (LTN), exponential smoothing (ES), median filtering (MF), gaussian window smoothing (GWS), z-score standardization (ZSS), local regression- weighted linear least squares and a first order polynomial model (LR1), local regression-weighted linear least squares and a second order polynomial model (LR2) and Savitzky-Golay filtering (SG) were used to preprocess the original spectral data (RD) (Ruszczak & Boguszewska-Mańkowska, 2022; Zou et al., 2022).

2.5 Predictive model

The data set input to the model, the columns are 256 spectral channels (spectral bands), and the rows are the spectral reflectance intensity of each column channel. 70% of the spectral curve data is randomly selected as the training set, and the remaining 30% of the spectral curve data is used as the test set. Use four supervised machine learning algorithms. Use EXtreme Gradient Boosting (XGBoost), Gradient Boosting Categorical Features (CatBoost), Light Gradient Boosting Machine (LightGBM), Stacking integrated (Stacking) machine learning algorithm to build a model to predict the water content of potatoes. These four classes of algorithms perform well in classification models and are widely used in other applications.

XGBoost algorithm

The XGBoost algorithm has high prediction accuracy and it was developed by modifying and improving the integrated tree model and the gradient boosting tree model. One of the advantages of this algorithm is that it avoids overfitting (Ji et al., 2019; Pan et al., 2022; Sun et al., 2021).

The integrated model of the tree is represented by Equation 3.

$$\hat{y}_i = \sum_{K=1}^{K} f_K(x_i), f_k \in F \tag{3}$$

In Equation 3, $\hat{y}_i$ denotes the predicted value, K is the number of trees, F is the collection space, x_i is the feature vector of the first data point, T is the number of leaves on the tree, and f_k is related to the k-th independent structure (q) and leaf weight (w). The XGBoost model loss function (Equation 3) comprises two components, i.e., classification and regression loss.

Equation 4:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{4}$$

In Equation 4, $\sum_{i=1}^{n} l(y_i, \bar{y}_i)$ denotes the training error between the predicted and observed values, and $\sum_{k=1}^{K} \Omega(f_k)$ denotes the sum of the complexity of the tree, which is a regular term used to control the complexity of the model (Equation 5).

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2 \tag{5}$$

In Equation 5, γ and λ represent the penalty coefficients.

The model performs better when the loss function is smaller. A greedy algorithm is used to divide the subtree and enumerate the feasible segmentation points, where the maximum gain obtained is calculated each time a new segment is added to an existing leaf (Chen & Guestrin, 2016). The gain is calculated with Equation 11. Equation 6:

$$Gain \cong \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)}{H_R + H_L + \lambda} - \gamma\right] \tag{6}$$

In Equation 6, the first and second terms represent the gains generated after splitting the left and right subtrees, respectively, and the third term is the gain without subtree splitting.

Catboost algorithm

The CatBoost algorithm is a gradient boosting decision tree (GBDT) algorithm that processes categorical data by performing stochastic permutations. This algorithm effectively prevents overfitting by conducting multiple permutations to train different models, thereby obtaining unbiased estimates of the gradients with little impact on the gradient estimation bias. The robustness of the model is high (Huang et al., 2019). CatBoost is an effective method for converting categorical data into numeric data and preventing overfitting. Categorical data are mainly preprocessed according to the following three steps (Hancock & Khoshgoftaar, 2020; Samat et al., 2021; Zhang et al., 2021).

1. Randomly arrange the initial data to generate multiple random arrangements.

2. Convert the tag value comprising a floating point or category into an integer.

3. Convert a categorical variable into a numeric variable using Equation 7.

$$avg_{target} = \frac{countInClass + prior}{totalCount + 1} \tag{7}$$

In Equation 12, *countInClass* denotes the frequency of the object tag with the current classification eigenvalue of 1, *totalCount* is the total number of objects with the classification eigenvalue that matches the current value, and *prior* is the initial value of the numerator. Advantages of the CatBoost algorithm include the capacity to handle categorical and numerical variables, support for customized loss functions, and obtaining accurate predictions in a low simulation time even when using the default parameters.

LightGBM algorithm

LightGBM is a novel GBDT algorithm that has been used widely in various data mining projects and competitions. The LightGBM algorithm includes two new techniques that involve gradient-based unilateral sampling and exclusive feature bundling (Sun et al., 2020). Based on the supervised training set $X = \{(x_i, y_i)\}_{i=1}^{n}$, LightGBM aims to find an approximation $\hat{y}(x)$ of a specific function $f^*(x)$ by minimizing the specific loss function $L(y, f(x))$. Equation 8:

$$\hat{y} = argminE_{y,X} L(y, f(x)) \tag{8}$$

The regression tree can be expressed as $^{w}q(x)$, $q \in \{1, 2, \ldots, J\}$, where J represents the number of leaves, q represents the decision rule for the tree sum, and w represents a vector leaf node,w is the sample weight of a vector leaf node (Dev & Eden, 2019). Thus, LightGBM can be trained in addition form at step t, as shown in Equation 9:

$$\Gamma_t = \sum_{i=1}^{n} L(y_i, F_{t-1}(x_j) + f_t(x_i)) \tag{9}$$

In Equation 10, Γ_T^* is a scoring function q for measuring the quality of the tree structure. The increased objective function obtained after splitting is shown in Equation 11.

$$\Gamma_T^* = -\frac{1}{2}\sum_{j=1}^{J} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} \tag{10}$$

$$G = \frac{1}{2}\left(\frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in R} h_i + \lambda} + \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda}\right) \tag{11}$$

In Equation 11, I_L and I_R denote the sample sets for the left and right branches, respectively. LightGBM grows the tree vertically whereas other algorithms such as *XGBoost* and GBDT grow it horizontally (level-wise growth) (Zhang et al., 2019). Vertical growth is more prone to overfitting, so LightGBM is an effective alternative algorithm only for large data sets. The accuracy of the predictions is often affected significantly by the hyperparameters. Thus, it is necessary to identify the number and range of the hyperparameters before using LightGBM.

Stacking integrated algorithm

The Stacking machine learning framework generalizes the output values from multiple models to improve the overall prediction performance (Figure 3). When using the Stacking integrated algorithm, the original data set is divided into several subdata sets, which are then employed as the input data for different base learners in the first layer (Shi & Zhang, 2019). The prediction values derived from the first layer are employed as the input data for the second layer to train the base learners.

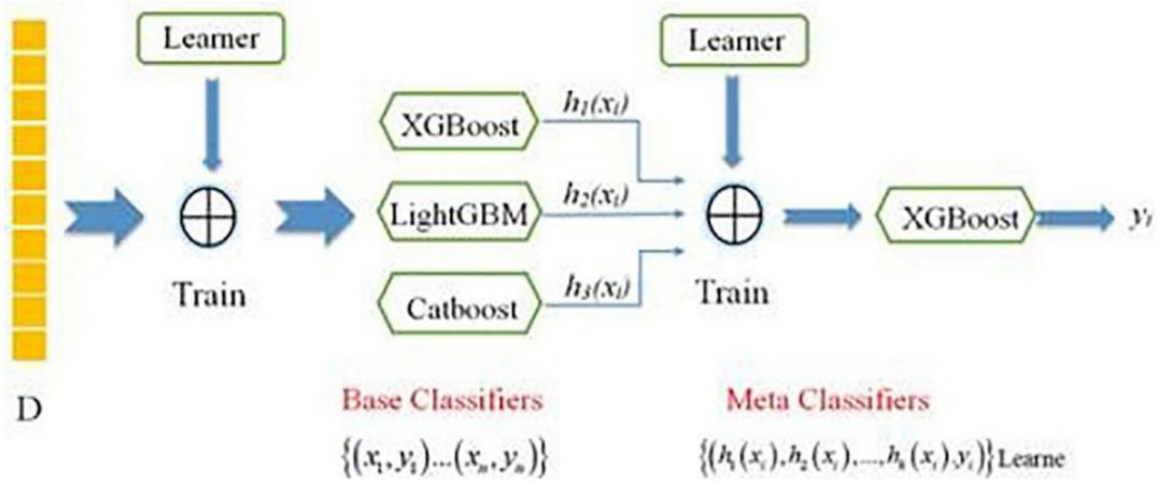

Figure 3. Schematic illustration of the Stacking machine learning framework.

The final prediction values are derived from the model of the second layer.

2.6 Hyperspectral feature selection and modeling

Hyperspectral and high-dimensional variables often contain large amounts of irrelevant information and redundant variables will affect the accuracy of the predictions produced by the final model. In the present study, CatBoost (Huang et al., 2019), Ridge algorithm (Piepho, 2009), LightGBM (Dev & Eden, 2019), LR (Friedman et al., 2010), Lasso (Tibshirani, 2011), XGBoost (Ji et al., 2019; Pan et al., 2022; Sun et al., 2021), Plsr (Long et al., 2019) and Stacking were used to effectively screen appropriate hyperspectral feature variables and to analyze the models.

2.7 Model performance assessment

A random sampling method was employed by using 75% of the sample reflection spectrum data as the training data and the remainder as the testing set. Different machine learning algorithms were employed to derive models for predicting the moisture contents of tuber by using the variables extracted from the hyperspectral images. The predictive performance was assessed using the mean absolute error (*Mae*), median absolute error (*Mdae*), root mean squared error (*Rmse*), coefficient of determination (R^2), and Fit_{Time}. The evaluation indicators used for the cross-validation set were *Rmse* and R^2. R^2 represent the proportion of the variance in the observed values that can be explained by that in the predicted values. *Rmse* measures the deviation between the observed and true values. *Mae* is the mean of the absolute error and it denotes the error of the predicted values. *Mdae* is calculated as the loss relative to the median value for all the absolute differences between the observed and predicted values, and provides a measure of the robustness of the variances (An et al., 2022; Liu et al., 2022; Pham & Liou, 2022). Equations 12-17:

$$\mathrm{Mae}\left(y_{\mathrm{pred}}, y_{\mathrm{act}}\right) = \frac{\sum_{i=1}^{n}\left|y_{\mathrm{pred}(n)} - y_{\mathrm{act}(n)}\right|}{n} \quad (12)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{\mathrm{pred}} - y_{\mathrm{act}}\right)^2}{\sum_{i=1}^{n}\left(y_{\mathrm{pred}} - y_{\mathrm{mean}}\right)^2} \quad (13)$$

$$R_{cv}^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{\mathrm{pred}} - y_{\mathrm{act}}\right)^2}{\sum_{i=1}^{n}\left(y_{\mathrm{pred}} - y_{\mathrm{mean}}\right)^2} \quad (14)$$

$$\mathrm{Rmse}_{cv} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{\mathrm{pred}} - y_{\mathrm{act}}\right)}{n}} \quad (15w)$$

$$\mathrm{Mdae}\left(y_{\mathrm{pred}}, y_{\mathrm{act}}\right) = \mathrm{median}\left(\left|y_{\mathrm{pred}(1)} - y_{\mathrm{act}(1)}\right|, \ldots, \left|y_{\mathrm{pred}(n)} - y_{\mathrm{act}(n)}\right|\right) \quad (16)$$

$$\mathrm{Rmse} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{\mathrm{pred}} - y_{\mathrm{act}}\right)}{n}} \quad (17)$$

In Equations 12-17, n denotes the number of samples, y_{act} is the observed value, y_{pred} is the predicted value, and y_{mean} is the mean of the measured values.

3 Results and discussion

3.1 Spectral reflectance pattern

Figure 4 shows the mean spectral reflectance values obtained for the time series during the oven-drying period. The reflection intensity of 500-900 nm potatoes decreased with the increase of baking time. This shows that this interval has a great relationship with the change of potato water content. In addition, and the four curves are obviously different, the reflectance value before dehydration is higher than the reflectance value after dehydration. The spectral reflectance curves obtained at the four sampling

points near the visible light wavelength of 400 nm basically agreed, with the minimum reflectance near the visible light wavelength of 450 nm where a trough formed. The reflectance increased in a linear manner in the visible light band from 500-900 nm. For each sample, the reflectance values decreased during oven drying and the four curves were clearly distinct. The reflectance values were higher before dehydration than after dehydration. A peak occurred close to the near infrared light wavelength at 930 nm and a trough at the near infrared band of 940-980 nm, where the reflectance value decreased. At wavelengths > 980 nm, the reflectance values increased and the four curves basically coincided. The curve declined and then increased sharply near the visible wavelength of 400 nm, mainly because this band contains the strong absorption bands for chlorophyll a and b, and it was affected by the electronic transition. The decrease in the reflectance value near the wavelength of 960 nm was due to the dominance of this band by the strong absorption band attributable to water and it was affected by C-H bond stretching and the first harmonic. The differences in the spectral reflectance curves suggest that hyperspectral imaging can be used to estimate the moisture contents of potatoes.

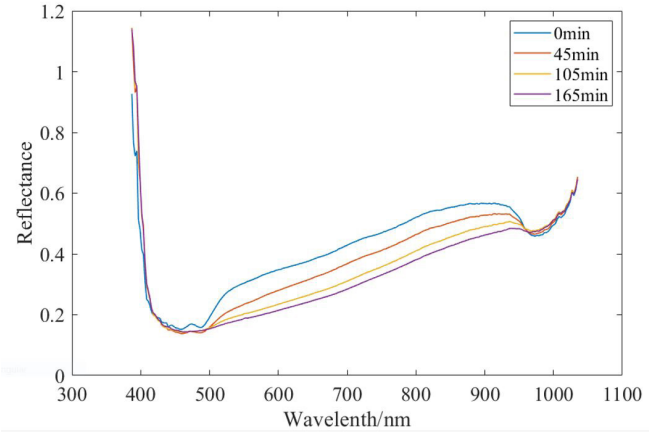

Figure 4. Spectral reflectance measured at four different sampling points.

3.2 Reflection spectrum preprocessing

The acquisition of spectral data is affected by the instrument's working status, detection environment, and the intensity of light. Noisy data may obscure the actual characteristics of the reflectance curves. Therefore, 17 data preprocessing methods. The processed data and raw data were trained with the CatBoost model, and the prediction results are shown in Table 1. As shown in Table 1, the FD algorithm had the largest *Mae* value, the ES algorithm had the smallest value, and the MF algorithm had the third lowest value. The FD algorithm produced the largest *Mdae* value whereas the MF algorithm had the smallest value. The FD algorithm yielded the largest *Rmse* value whereas the LR2 algorithm produced the smallest value. The MF algorithm yielded the largest R^2 value whereas the SD algorithm obtained the smallest. The SD algorithm obtained the largest *Rmse* value whereas the MF algorithm produced the smallest. The MF algorithm obtained the largest R^2_{cv} value whereas the SD algorithm produced the smallest. The L2NN algorithm yielded the largest Fit_{Time} value whereas the ES algorithm had the smallest. The prediction performance of the LR1 and SG algorithms did not differ significantly from that obtained by using the original data for modeling. The *Rmse* value was 0.0677 for both the LR1 and RD algorithms. The SG and RD algorithms obtained similar R^2 values around 0.6980. The results produced by the SG algorithm suggested that it was the most stable method, with relatively small differences in the

Table 1. Prediction results obtained with different pretreatments.

Methods	Mae	Mdae	Rmse	R^2	$Rmse_{cv}$	R^2_{cv}	Fit time
RD	0.0534	0.0452	0.0677	0.6980	0.0739	0.6914	23.8276
FD	0.0786	0.0581	0.1075	0.7203	0.0891	0.6876	29.4692
SD	0.0752	0.0626	0.0983	0.6574	0.0984	0.5238	24.8935
GWS	0.0590	0.0423	0.0748	0.6746	0.0747	0.7108	24.3578
BS	0.0577	0.0468	0.0749	0.7675	0.0717	0.6966	28.2548
LR1	0.0551	0.0509	0.0677	0.7173	0.0769	0.6753	28.5188
LR2	0.0510	0.0396	0.0659	0.7062	0.0783	0.7009	29.9864
L2NN	0.0555	0.0545	0.0707	0.6685	0.0769	0.6541	30.1779
LTN	0.0563	0.0417	0.0738	0.7354	0.0741	0.7371	28.1683
MAM	0.0618	0.0533	0.0772	0.6803	0.0708	0.7282	28.2339
MSC	0.0605	0.0490	0.0778	0.6656	0.0781	0.6393	27.2989
SG	0.0591	0.0517	0.0761	0.6991	0.0734	0.6833	22.3748
ACN	0.0575	0.0429	0.0779	0.6871	0.0745	0.7099	23.2816
WTD	0.0600	0.0521	0.0743	0.7235	0.0767	0.6770	25.7103
ES	0.0523	0.0437	0.0663	0.7321	0.0745	0.7194	22.3437
MF	0.0537	0.0393	0.0720	0.8015	0.0599	0.8021	23.4790
ZSS	0.0562	0.0440	0.0764	0.7525	0.0734	0.7075	24.4995
MMS	0.0598	0.0449	0.0809	0.7180	0.0775	0.7007	24.4379

The mae is the squared absolute error. The mdae is the median of the absolute error values between the actual and predicted values. The Rmse is root mean square error. The R^2 is the coefficient of absolute certainty. $Rmse_{cv}$ is root mean square error of cross validation. The R^2_{cv} is the coefficient of determination for cross-validation.

Rmse and R^2 values between the prediction and cross-validation data sets. However, the prediction accuracies with FD, SD, and MMS were lower than those when using the original data for modeling, where larger *Rmse* and smaller R^2 values were obtained. The differences in the prediction abilities when using data preprocessed with GWS, L2NN, MAM, MSC, and CAN compared with those using the raw data were relatively small, although *Rmse* was larger and R^2 was smaller for the models derived based on the data preprocessed by the LR2, ES, and MF algorithms relative to the models derived with the raw data, where the indicators obtained were comparable. The accuracy of the predictions obtained by models derived using the preprocessed data was much better than that based on the raw data because the raw data-based models obtained consistently low *Rmse* and larger R^2 values. In particular, the prediction models based on the data preprocessed with the MF algorithm had the smallest *Mdae* (0.0393) and highest R^2 values among all of the preprocessed data-based models. The lowest $Rmse_{CV}$ value was 0.0599 with the MF algorithm. The R^2 and R^2_{CV} values were comparable and they were the largest, thereby indicating that the MF algorithm exhibited high stability. The accuracies of the models derived using raw data and data preprocessed by the BS, LTN, and WTN algorithms were difficult to compare because all of the *Rmse* and R^2 values were larger for the models obtained with the preprocessed data compared with those based on the raw data. Therefore, it's concluded that the different data preprocessing methods had various effects on the accuracy of the model predictions, with more accurate predictions, inferior predictions, and some comparable predictions. Finally, after comparing the different model performance indicators, the MF algorithm was selected as the optimal algorithm.

3.3 Feature band extraction

The original hyperspectral curves obtained from images of potatoes contained 256 characteristic variables. However, these high-dimensional variables contained large amounts of irrelevant information and the redundant variables affected the classification accuracy for the final model. Therefore, multiple LR methods as well as the Ridge, Lasso, XGBoost, LightGBM, Plsr, CatBoost, and XGBoost algorithms were used to extract the feature wavelengths from the hyperspectral curves. The top 40 feature variables that made the greatest contributions to the hyperspectral curves were extracted and the weighted characteristic wavelengths were obtained (Table 2). The feature extraction algorithms, i.e., Plsr, CatBoost, Ridge, and Lasso, all identified the band around 850-970 nm as the primary feature variable. Using the LR algorithm, the weighted contribution of this band ranked fifth, where the peak occurred near a wavelength of 940 nm, and valleys and absorption characteristics were identified around 960 nm. Using the LightGBM and XGBoost algorithms, the band around 400 nm was the primary feature variable. With the Lasso algorithm, the weighted contribution of this band ranked sixth, and the valley and absorption characteristics were identified at 400 nm. Using the CatBoost algorithm, the contribution of the band at 694-865 nm ranked from fourth to eighth. The features with wavelengths of 833.66 nm and 917.5 nm ranked fourth and fifth, respectively, with the Lasso algorithm. The band at 547-848 nm ranked among the top 40 important features with LR. The wavelength of 957.34 nm ranked first using the Plsr algorithm. The band at 510-884 nm ranked in the top 40 with the Ridge algorithm. The spectral image was clearly distinguished in the band from 510-958 nm in the spectral curve, thereby verifying the importance of the extracted features. The experimental results showed that the feature extraction algorithms could extract common features at the same wavelengths, but significant differences were also found. The feature variables were used subsequently as input variables for the prediction algorithms.

3.4 Visualization of top five feature correlation coefficients for each model

Figure 5 shows the feature correlation coefficients with particular importance for each model. The selected features that had relatively strong correlations with the moisture contents are shown in red, blue, and purple in Figure 5. A small number of the features had relatively weak correlations and they are shown in green. The results suggested that the feature variables

Table 2. The top 40 characteristic bands selected by different algorithms.

Models	Feature band (nm)
CatBoost	579.93 997.38 403.88 694.86 735.62 864.98 697.4 449.59 604.3 975.99 831.06 447.18 1021.53 522.66 1002.73 396.7 439. 93 661.98 727.95 387.15 399.09 468.97 391.92 768.96 748.42 740.75 797.34 459.27 601.81 1024.22 495.27 761.25 413.47 515.31 1000.05 552.2 928.15 805.1 1032.29 490.87
Lasso	957.34 949.36 970.65 833.66 917.58 478.69 435.11 684.72 1002.73 1008.1 1032.29 554.67 396.7 387.15 1024.22 468.97 1010.78 427.89 510.41 432.7 488.43 473.83 1000.05 401.49 454.43 1034.99 415.87 1018.84 408.67 1026.91 423.07 997.38 456.65 391.92 420.67 439.93 413.47 464.12 1016.15 389.54
LightGBM	396.7 387.15 408.67 391.92 389.54 399.09 415.87 394.31 401.49 418.27 403.88 406.28 413.47 1032.29 1029.6 411.07 466.55 1021.22 425.48 442.35 423.07 503.07 1034.99 420.67 439.93 986.67 498.19 1010.78 471.4 435.11 481.12 456.85 500.63 1021.53 437.52 981.33 1018.84 1026.91 452.01 477.18
Linear	705.02 659.46 557.14 684.72 823.26 862.36 682.19 559.61 694.86 661.92 825.85 576.94 654.42 699.94 692.33 601.81 651.9 717.75 571.98 712.66 626.79 756.11 687.26 656.94 569.5 878.08 547.26 697.4 774.11 649.39 789.58 619.28 564.55 584.39 743.3 586.87 672.08 756.68 636.84 634.31
Plsr	957.34 970.65 435.11 387.15 1026.91 1024.22 1008.1 1010.78 408.67 396.7 949.36 967.99 391.92 997.38 1029.6 473.83 478.69 962.66 1018.84 468.97 914.93 1000.05 464.12 1034.99 687.26 656.94 569.5 878.08 909.66 439.93 917.58 488.43 413.47 437.52 423.07 920.22 510.41 928.15 912.29 493.3
Ridge	854.52 604.3 833.66 584.39 510.41 789.58 815.47 781.84 893.33 818.06 684.86 766.39 802.51 851.91 571.98 629.3 812.87 614.28 544.8 616.78 624.29 875.46 567.03 507.96 594.33 532.49 542.33 609.29 805.1 771.54 702.48 779.26 768.96 554.67 539.87 1021.53 522.66 712.66 730.51 667.03
XGBoost	387.15 389.54 411.07 391.92 415.87 399.09 413.47 396.7 394.34 403.88 498.18 406.28 473.83 401.49 418.27 425.48 439.93 420.67 510.41 408.67 423.07 500.63 471.4 503.07 495.74 430.29 435.11 507.96 1034.99 427.89 412.86 442.35 537.41 444.76 476.26 452.01 505.52 515.31 454.43 712.66

extracted by the supervised algorithm were highly correlated (red and orange in Figure 5. Thus, the main features selected were reasonably consistent. Most of the features were strongly and positively correlated, thereby suggesting that there was redundancy between the features. Therefore, the best prediction would not be produced by combining the features extracted from several models. It was reasonable and feasible to use the feature extracted from each model separately for modeling.

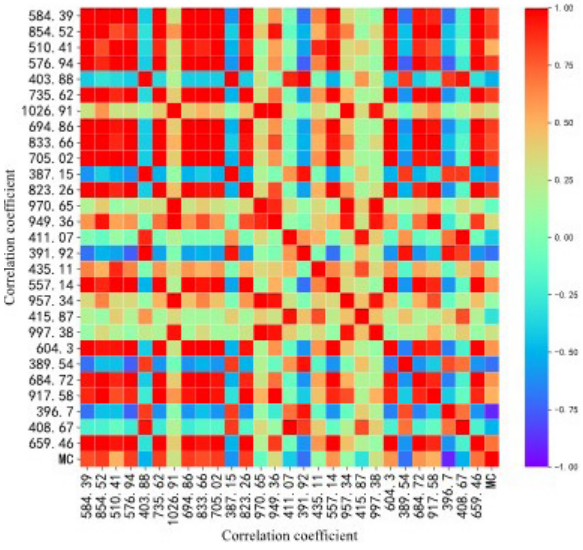

Figure 5. Correlation coefficient map showing the top five features for each mode.

3.5 Analysis of hyperspectral response to the moisture content of potatoes

Analysis of hyperspectral response to the moisture content of potatoes

The results obtained from the four prediction models by using the feature wavelengths selected with different algorithms are summarized in Table 3. The Stacking model was establishd by taking XGBoost, CatBoost, and LightGBM as the first layer, and the XGBoost model as the second layer. As shown in Table 3, the Stacking model explained > 80% of the variance regardless of the feature extraction method employed, whereas some other models only explained < 80% of the variance in the moisture contents of potato tubers, and there were large variations in the prediction abilities of the models. Thus, the Stacking model performed best overall. In addition, the XGBoost model required the least time whereas the Stacking model consumed the most time. The models produced using feature extraction methods performed better

Table 3. Comparison of the indexes obtained for the models established using the feature wavelengths extracted by different algorithms.

Models	Methods	Mae	Mdae	Rmse	R^2	$Rmse_{cv}$	R^2_{cv}	Fit time
XGBoost	CatBoost	0.0485	0.0346	0.0679	0.8187	0.0559	0.8218	0.6333
	Lasso	0.0477	0.0389	0.0610	0.8908	0.0544	0.8448	0.5056
	LightGBM	0.0472	0.0332	0.0620	0.8764	0.0592	0.8076	0.6034
	Linear	0.0457	0.0291	0.0616	0.7842	0.0703	0.7442	0.4618
	Plsr	0.0446	0.0357	0.0554	0.8578	0.0545	0.8338	0.7201
	Ridge	0.0505	0.0409	0.0647	0.8244	0.0701	0.7285	1.0901
	XGBoost	0.0388	0.0298	0.0516	0.8795	0.0551	0.8343	0.6971
LightGBM	CatBoost	0.0406	0.0359	0.0511	0.8482	0.0537	0.8189	1.3105
	Lasso	0.0408	0.0321	0.0524	0.8414	0.0546	0.8158	1.4667
	LightGBM	0.0410	0.0308	0.0541	0.8576	0.0577	0.7938	1.3045
	Linear	0.0462	0.0344	0.0614	0.7288	0.0640	0.7365	0.8846
	Plsr	0.0413	0.0330	0.0525	0.8443	0.0524	0.8235	1.5148
	Ridge	0.0460	0.0345	0.0588	0.7396	0.0635	0.7383	1.5437
	XGBoost	0.0408	0.0323	0.0529	0.8511	0.0542	0.8199	1.4595
CatBoost	CatBoost	0.0594	0.0431	0.0798	0.8187	0.0622	0.8097	9.7620
	Lasso	0.0553	0.0415	0.0717	0.8145	0.0636	0.7947	10.8851
	LightGBM	0.0560	0.0419	0.0720	0.8014	0.0636	0.7469	10.2706
	Linear	0.0569	0.0478	0.0707	0.6495	0.0746	0.6302	8.6289
	Plsr	0.0509	0.0400	0.0648	0.7965	0.0580	0.8217	11.4537
	Ridge	0.0495	0.0389	0.0647	0.7018	0.0655	0.6962	9.6833
	XGBoost	0.0557	0.0468	0.0717	0.8121	0.0646	0.7979	11.7052
Stacking	CatBoost	0.0439	0.0380	0.0586	0.8561	0.0543	0.8211	14.0775
	Lasso	0.0428	0.0356	0.0540	0.8690	0.0545	0.8351	14.1991
	LightGBM	0.0485	0.0344	0.0667	0.8679	0.0582	0.8075	13.0800
	Linear	0.0466	0.0350	0.0628	0.8113	0.0739	0.7395	10.0741
	Plsr	0.0410	0.0346	0.0522	0.8452	0.0554	0.8255	14.2234
	Ridge	0.0489	0.0381	0.0615	0.8076	0.0695	0.7353	13.6310
	XGBoost	0.0405	0.0316	0.0526	0.8640	0.0563	0.8169	17.6071

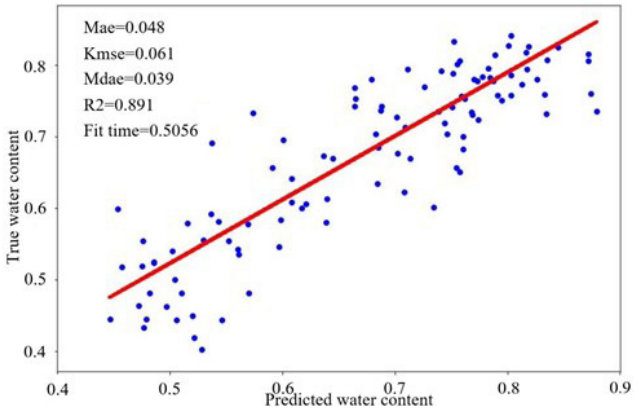

Figure 6. Curve fitting diagram of predicted value and actual value.

compared with the full-band CatBoost predictions. Among the seven feature extraction methods, the features extracted using the LR model followed by Ridge were the worst in terms of the modeling and prediction abilities compared with the other methods. The feature variables extracted by XGBoost and Lasso were much better in terms of the modeling performance. As shown in Figure 6, best model produced by combining the Lasso and XGBoost algorithms achieved an *Rmse* of 0.061, R^2 of 0.8908, *Mdae* of 0.0389, and $R^2_{cv}$of 0.8448. The worst model generated by combining the LR + CatBoost algorithms had an *Rmse* of 0.0707, R^2 of 0.6495, *Mdae* of 0.0478, and R^2_{cv} of 0.630.

4 Conclusion

In this article, potato tubers were heated at 120 degrees Celsius for 0 min, 45 min, 105 min and 165 min. Through spectral data preprocessing, feature extraction and data modeling, the prediction of moisture content of potato tubers under different baking times was successfully achieved. In this study, 17 hyperspectral data preprocessing methods were used to eliminate the effect of noise in the original hyperspectral data of potato images. The preprocessing results show that using MF preprocessing is better than other methods in terms of its predictive performance. Therefore, choosing an appropriate preprocessing method can reduce the noise and improve the prediction accuracy of the model. In addition, spectral data in the 400 nm and 547-970 nm bands are important for predicting the moisture content of potato tubers. Among them, the top forty feature bands extracted by lasso make a significant contribution to the prediction accuracy of the model. The best model generated by combining Lasso and XGBoost algorithms has Rmse of 0.0610, R^2 of 0.8908, *Mdae* of 0.0389, and R^2_{cv} of 0.8448. To sum up, the best prediction model is MF-Lasso-XGBoost. Using hyperspectral imaging technology can accurately predict the water content of potato tubers. At the same time, it can also provide new opportunities for future crop moisture detection related ideas.

Conflict of interest

The authors declare no conflict of interest.

Availability of data and material

Samples of the compounds are not available from the authors.

Funding

This study was supported by the National Natural Science Foundation of China (31601227, 31501221, 61803325). The editors and anonymous reviewers are thanked for their help with improving the quality of this article.

References

An, T., Huang, W., Tian, X., Fan, S., Duan, D., Dong, C., Zhao, C., & Li, G. (2022). Hyperspectral imaging technology coupled with human sensory information to evaluate the fermentation degree of black tea. *Sensors and Actuators B: Chemical*, 366, 131994. http://dx.doi.org/10.1016/j.snb.2022.131994.

Calderón, L. A., Mena, V. A. G., & Miranda, J. M. R. (2021). Development of an extruded food product similar to fried potatoes, based on by-products of potatoes and rice. physicochemical and microbiological evaluation. *Food Science and Technology*, 41(2), 359-364. http://dx.doi.org/10.1590/fst.03820.

Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In B. Krishnapuram, & M. Shah (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). New York: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Das, B., Manohara, K. K., Mahajan, G. R., & Sahoo, R. N. (2020). Spectroscopy based novel spectral indices, PCA- and PLSR-coupled machine learning models for salinity stress phenotyping of rice. *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy*, 229, 117983. http://dx.doi.org/10.1016/j.saa.2019.117983. PMid:31896051.

Dev, V. A., & Eden, M. R. (2019). Formation lithology classification using scalable gradient boosted decision trees. *Computers & Chemical Engineering*, 128, 392-404. http://dx.doi.org/10.1016/j.compchemeng.2019.06.001.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1-22. http://dx.doi.org/10.18637/jss.v033.i01. PMid:20808728.

Gerhards, M., Rock, G., Schlerf, M., & Udelhoven, T. (2016). Water stress detection in potato plants using leaf temperature, emissivity, and reflectance. *International Journal of Applied Earth Observation and Geoinformation*, 53, 27-39. http://dx.doi.org/10.1016/j.jag.2016.08.004.

Habig, J. W., Rowland, A., Pence, M. G., & Zhong, C. X. (2018). Food safety evaluation for R-proteins introduced by biotechnology: a case study of VNT1 in late blight protected potatoes. *Regulatory Toxicology and Pharmacology*, 95, 66-74. http://dx.doi.org/10.1016/j.yrtph.2018.03.008. PMid:29530614.

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, 7(1), 94. http://dx.doi.org/10.1186/s40537-020-00369-8. PMid:33169094.

Hou, Y., Zhao, P., Zhang, F., Yang, S., Rady, A., Wijewardane, N. K., Huang, J., & Li, M. (2022). Fourier-transform infrared spectroscopy and machine learning to predict amino acid content of nine commercial insects. *Food Science and Technology*, 42, e100821. http://dx.doi.org/10.1590/fst.100821.

Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., Zeng, W., & Zhou, H. (2019). Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *Journal of Hydrology*, 574, 1029-1041. http://dx.doi.org/10.1016/j.jhydrol.2019.04.085.

Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An application of a three-stage XGBoost-based model to sales forecasting of a cross-border

e-commerce enterprise. *Mathematical Problems in Engineering*, 2019, 8503252. http://dx.doi.org/10.1155/2019/8503252.

Liu, N., Wu, L., Chen, L., Sun, H., Dong, Q., & Wu, J. (2018). Spectral characteristics analysis and water content detection of potato plants leaves. *IFAC-PapersOnLine*, 51(17), 541-546. http://dx.doi.org/10.1016/j.ifacol.2018.08.152.

Liu, Y., Zhou, S., Wu, H., Han, W., Li, C., & Chen, H. (2022). Joint optimization of autoencoder and self-supervised classifier: anomaly detection of strawberries using hyperspectral imaging. *Computers and Electronics in Agriculture*, 198, 107007. http://dx.doi.org/10.1016/j.compag.2022.107007.

Long, Z., Wang, Y., Liu, X., & Yao, L. (2019). Two-step paretial least square regression classifiers in brain-state decoding using functional magnetic resonance imaging. *PLoS One*, 14(4), e0214937. http://dx.doi.org/10.1371/journal.pone.0214937. PMid:30970029.

Nikzad, N., Ghavami, M., Seyedain-Ardabili, M., Akbari-Adergani, B., & Azizinezhad, R. (2021). Effect of deep frying process using sesame oil, canola and frying oil on the level of bioactive compounds in onion and potato and assessment of their antioxidant activity. *Food Science and Technology*, 41(3), 545-555. http://dx.doi.org/10.1590/fst.35819.

Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science Engineering*, 208, 109520. http://dx.doi.org/10.1016/j.petrol.2021.109520.

Pereira, A. M., Petrucci, K. P. O. S., Gomes, M. P., Gonçalves, D. N., Cruz, R. R. P., Ribeiro, F. C. S., & Finger, F. L. (2021). Quality of potato CV. innovator submitted refrigeration and recondition. *Food Science and Technology*, 41(1), 34-38. http://dx.doi.org/10.1590/fst.26619.

Pham, Q. T., & Liou, N.-S. (2022). The development of on-line surface defect detection system for jujubes based on hyperspectral images. *Computers and Electronics in Agriculture*, 194, 106743. http://dx.doi.org/10.1016/j.compag.2022.106743.

Piepho, H. P. (2009). Ridge regression and extensions for genomewide selection in maize. *Crop Science*, 49(4), 1165-1176. http://dx.doi.org/10.2135/cropsci2008.10.0595.

Ruszczak, B., & Boguszewska-Mańkowska, D. (2022). Deep potato – the hyperspectral imagery of potato cultivation with reference agronomic measurements dataset: towards potato physiological features modeling. *Data in Brief*, 42, 108087. http://dx.doi.org/10.1016/j.dib.2022.108087. PMid:35392624.

Samat, A., Li, E., Du, P., Liu, S., & Xia, J. (2021). GPU-accelerated CatBoost-forest for hyperspectral image classification via parallelized mRMR ensemble subspace feature selection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 3200-3214. http://dx.doi.org/10.1109/JSTARS.2021.3063507.

Santos, M. N. S., Lima, P. C. C., Araujo, F. F., Araújo, N. O., & Finger, F. L. (2020). Activity of polyphenoloxidase and peroxidase in non-dormant potato tubers treated with sprout suppressors. *Food Science and Technology*, 40(suppl. 1), 222-227. http://dx.doi.org/10.1590/fst.08119.

Shao, Y., Shi, Y., Qin, Y., Xuan, G., Li, J., Li, Q., Yang, F., & Hu, Z. (2022). A new quantitative index for the assessment of tomato quality using Vis-NIR hyperspectral imaging. *Food Chemistry*, 386, 132864. http://dx.doi.org/10.1016/j.foodchem.2022.132864. PMid:35509167.

Shi, J., & Zhang, J. (2019). Load forecasting based on multi-model by stacking ensemble learning. *Chinese Society for Electrical Engineering*, 39(14), 4032-4042. https://doi.org/10.13334/j.0258-8013.pcsee.181510.

Su, W. H., & Sun, D. W. (2016). Potential of hyperspectral imaging for visual authentication of sliced organic potatoes from potato and sweet potato tubers and rapid grading of the tubers according to moisture proportion. *Computers and Electronics in Agriculture*, 125, 113-124. http://dx.doi.org/10.1016/j.compag.2016.04.034.

Sun, B., Sun, T., & Jiao, P. (2021). Spatio-temporal segmented traffic flow prediction with ANPRS data based on improved XGBoost. *Journal of Advanced Transportation*, 2021(Spe), 5559562. http://dx.doi.org/10.1155/2021/5559562.

Sun, H., Chen, X., Sun, Z., Minzan, L. I., Zhang, M., & Jingzhu, W. U. (2018). Rapid detection of moisture content in maize leaves based on transmission spectrum. *Chinese Journal of Agricultural Engineering*, 49(03), 173-178.

Sun, H., Liu, N., Wu, L., Zheng, T., Li, M. Z., & Wu, J. Z. (2019). Visualization of water content distribution in potato leaves based on hyperspectral image. *Spectroscopy Spectral Analysis*, 39(3), 910-916.

Sun, X. L., Liu, M. X., & Sima, Z. Q. (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters*, 32, 101084. http://dx.doi.org/10.1016/j.frl.2018.12.032.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273-282. http://dx.doi.org/10.1111/j.1467-9868.2011.00771.x.

Wang, X. W., Xing, X. Y., Zhao, M. C., & Yang, J. R. (2021). Comparison of multispectral modeling of physiochemical attributes of greengage: Brix and pH values. *Food Science and Technology*, 41(suppl. 2), 611-618. http://dx.doi.org/10.1590/fst.21320.

Zhang, J., Mucs, D., Norinder, U., & Svensson, F. (2019). LightGBM: an effective and scalable algorithm for prediction of chemical toxicity-application to the Tox21 and mutagenicity data sets. *Journal of Chemical Information and Modeling*, 59(10), 4150-4158. http://dx.doi.org/10.1021/acs.jcim.9b00633. PMid:31560206.

Zhang, J., Zhang, D., Cai, Z., Wang, L., Wang, J., Sun, L., Fan, X., Shen, S., & Zhao, J. (2022). Spectral technology and multispectral imaging for estimating the photosynthetic pigments and SPAD of the Chinese cabbage based on machine learning. *Computers and Electronics in Agriculture*, 195, 106814. http://dx.doi.org/10.1016/j.compag.2022.106814.

Zhang, M., Chen, W., Zhang, Y., Liu, F., Yu, D., Zhang, C., & Gao, L. (2021). Fault diagnosis of oil-immersed power transformer based on difference-mutation brain storm optimized catboost model. *IEEE Access*, 9, 168767-168782. http://dx.doi.org/10.1109/ACCESS.2021.3135283.

Zhao, R., An, L., Song, D., Li, M., Qiao, L., Liu, N., & Sun, H. (2021). Detection of chlorophyll fluorescence parameters of potato leaves based on continuous wavelet transform and spectral analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 259, 119768. http://dx.doi.org/10.1016/j.saa.2021.119768. PMid:33971438.

Zheng, T., Liu, N., Wu, L., Li, M., Sun, H., Zhang, Q., & Wu, J. (2018). Estimation of chlorophyll content in potato leaves based on spectral red edge position. *IFAC-PapersOnLine*, 51(17), 602-606. http://dx.doi.org/10.1016/j.ifacol.2018.08.131.

Zhu, Y., Yuan, Y., Mei, L., Ding, S., Gao, Y., Du, X., & Guo, L. (2022). Comparison of structural and physicochemical properties of potato protein and potato flour modified with tyrosinase. *Journal of Integrative Agriculture*, 21(5), 1513-1524. http://dx.doi.org/10.1016/S2095-3119(21)63852-2.

Zou, Z., Wang, L., Chen, J., Long, T., Wu, Q., & Zhou, M. (2022). Research on peanut variety classification based on hyperspectral image. *Food Science and Technology*, 42, e18522. http://dx.doi.org/10.1590/fst.18522.