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

DOI: https://doi.org/10.1590/fst.71322

Nondestructive detection of peanuts mildew based on hyperspectral image technology and machine learning algorithm

Zhiyong ZOU[1], Jie CHEN[1], Li WANG[1], Weijia WU[1], Tingjiang YU[2], Yuchao WANG[1], Yongpeng ZHAO[1], Peng HUANG[1], Bi LIU[1], Man ZHOU[3], Ping LIN[4], Lijia XU[1]*

Abstract

In order to realize the rapid nondestructive detection of mildew peanut in the process of peanuts storage, hyperspectral imaging technology was proposed to detect mildew peanut. A total of 200 peanuts were selected from 5 kinds of peanuts purchased in the market for moldy treatment, and the remaining 400 peanuts were kept sterile. After completion, samples were collected with a hyperspectral instrument to obtain spectral data of the samples. According to the characteristics of the data, 10 pre-processing algorithms were used to de-noise the data, and Median Filtering (MF) had the best effect, with the recognition accuracy reaching 97.7%. GBDT, LightGBM, CatBoost and XGBoost algorithms were used to extract important feature bands in the spectral data pre-processed by MF. GBDT, LightGBM, CatBoost and XGBoost algorithms were used to model the extracted feature bands. The results showed that LightGBM is the best algorithm with a detection rate of 99.10%. Optuna algorithm was used to tune its parameters. Compared with the previous model, the running time of the optimized model was improved by about 0.25 s. The results showed that hyperspectral imaging provides an efficient and nondestructive method for detecting mildew in peanut storage.

Keywords: hyperspectral; mildew; optuna; LightGBM.

Practical Application: Application of hyperspectral in peanut mildew detection.

1 Introduction

With global peanuts production reaching 50.63 million tons in 2021, a large number of peanuts are facing storage problems. In the storage process, with the change of climate, peanuts are prone to mildew due to humidity. Aflatoxin poisons (AFT) naturally formed after mildew include aflatoxin poisons B1, B2, G1 and G2 (Idris et al., 2010; Mutua et al., 2019), which are toxic secondary metabolites formed by aflatoxin (Gonçalves et al., 2022) and parasitic aspergillus, among which AFB1 is the most commonly used toxic type. AFT is classified as a group I human carcinogen by the Global Organization for Scientific Research on Cancer (IARC) because it is toxic, teratogenic and genetically toxic to human and mammalian liver. In humans, AFB1 is metabolized by DNA-binding enzymes to form AFB1-DNA adducts, causing acute poisoning and increasing the risk of hepatocellular carcinoma (HCC) (Yang et al., 2019). Therefore, many countries and organizations, such as the Codex Alimentarius Commission (CAC), the European Union (EU), the United States (US), Japan and South Korea, have established maximum levels (MLs) of aflatoxin in peanuts and peanuts products ranging from 4.0 μg/kg to 20.0 μg/kg to reduce AFT exposure (Gonçalves et al., 2017). The current National Food Safety Standard (GB 2761-2017) stipulates that the ML of AFB1 in peanuts and peanut products is only 20.0 μg/kg (Zhou et al., 2016). Although various chemical, physical and biological methods have been applied to control AFT in food, AFT contamination is still a major food safety issue of global concern (Kumar et al., 2017). In China, peanuts are one of the major oil crops and high-consumption agricultural products, and severe AFT pollution in peanuts poses a threat to human health and trade (Sun et al., 2017). Therefore, it is important to realize the rapid, nondestructive and effective identification of moldy peanuts for the safety of peanut food processing.

In recent years, the use of hyperspectral image processing technology for non-destructive testing of agricultural products has not been developed for a long time but has made rapid progress, and corresponding research results have been obtained (Guo et al., 2019; Zhang et al., 2019). Yuan et al. (2020) researched the identification of peanuts mildew using a small number of critical wavelength bands and an integrated classifier based on hyperspectral images. Simulating the natural process of peanuts fungal infection, three experiments were conducted on peanuts with different varieties of growing mold, and detailed hyperspectral images of healthy and moldy peanuts were captured in the 960-2568 nm range. Based on the hyperspectral image, the peanuts spectral images were obtained by combining genetic algorithm and continuous algorithm, and the projection algorithm was used to select the key wavelengths. After that, EC was composed of support vector machine (SVM), partial least squares discriminant analysis was performed by clustering independent soft pattern

Received 02 June, 2022
Accepted 23 July, 2022

[1] *College of Mechanical and Electrical Engineering, Sichuan Agricultural University, Ya 'an, China*
[2] *State Energy Dadu River Waterfall Ditch Hydroelectric Power Plant, Ya 'an, China*
[3] *College of Food Science, Sichuan Agricultural University, Ya 'an, China*
[4] *School of Electrical Engineering and Automation, Hubei Normal University, Huangshi, China*
*Corresponding author: xulijia@sicau.edu.cn

classifier (SIMCA), according to the selected key wavelengths (982 nm, 1180 nm, 1405 nm, 1540 nm, 1871 nm, 1938 nm, 1999 nm). The overall classification accuracy of EC, SVM, PLS-DA and SIMCA at pixel level was 97.66%, 97.53%, 95.31% and 97.36%, respectively. Finally, the kernel scale classification diagram showed the effectiveness of the method and the distribution of moldy peanuts in healthy peanuts. This indicates that NIL-HSI is a reliable prediction method for analyzing peanut mildew (Liu et al., 2020). Liu et al. (2020) collected 16 hyperspectral images of healthy peanuts, damaged peanuts and moldy peanuts from 1066 peanut samples with a spectrometer. Deeplab V3+, Segnet and Unet Hypernet were constructed as control models for comparison. The proposed peanut recognition index (PRI) was fused to peanut recognition, in which hyperspectral images were used as pre-extraction of data features, and multi-feature fusion block (MF block) was constructed to be integrated into the control model as enhancement of model features to improve the accuracy of peanuts recognition (Borregaard et al., 2000; Yuan et al., 2020). Nakariyakul & Casasent (2011) developed a new method to detect internal damage of almonds, which only requires two sets of ratio features (response ratios of two different spectral bands are classified). The proposed method avoids the use of first-order search to thoroughly search the entire feature space. Firstly, the ratio feature set was sorted, and then the optimal ratio feature was selected according to the sorting set. It is feasible to use ratio feature to classify, which can be used in real time multispectral analysis sensor system. Experimental results showed that this method had a high classification rate, whether using the best feature of an independent band to select a subset or using feature extraction to use all wavelength data (Nakariyakul & Casasent, 2011).

Food microorganisms are affected by factors such as environment and time, and the detection cycle is long, so that the microbial detection data related to food quality and safety cannot relatively truly reflect the situation of food microorganisms. In recent years, near-infrared detection technology has attracted more and more attention in the field of food microbiological detection due to its high efficiency, non-destructiveness and rapidity. The spectrum of the sample has a certain correlation with the measured value of the composition and property parameters of the sample. The chemometrics method is used to correlate them, and the quantitative or qualitative relationship between the two is established. For biochemical detection, qualitative and quantitative analysis of microorganisms in the sample can be performed only by collecting the spectrum of the sample to be predicted.

Pranoto et al. (2022) used artificial neural network and fluorescence spectroscopy to identify food vegetable oils with an accuracy rate of 72% (Pranoto et al., 2022). Hou et al. (2022) used Fourier transform infrared spectroscopy and machine learning to predict the amino acid content of nine commercial insects with a coefficient of determination of 0.97 (Hou et al., 2022). Chen & Yu (2022) tested the food safety sampling inspection system based on deep learning, and applied various technical means such as image processing, speech recognition, and target detection (Chen & Yu, 2022). Using Differential Scanning Calorimetry (DSC) combined with machine learning tools to detect adulteration in raw milk, Farah et al. (2021) achieve 100% identification and predictive power (Farah et al., 2021).

2 Materials and methods

2.1 The experimental materials

In recent years, major planting provinces (Shandong, Henan, Jiangsu, etc.) the main peanut varieties (Dabaisha, Huayu 16, Xiaobaisha, Haihua, Luhua) as the research object, as shown in Figure 1, and in each selected peanuts production phase is full of 120, a total of 600, each of the randomly selected 40 make

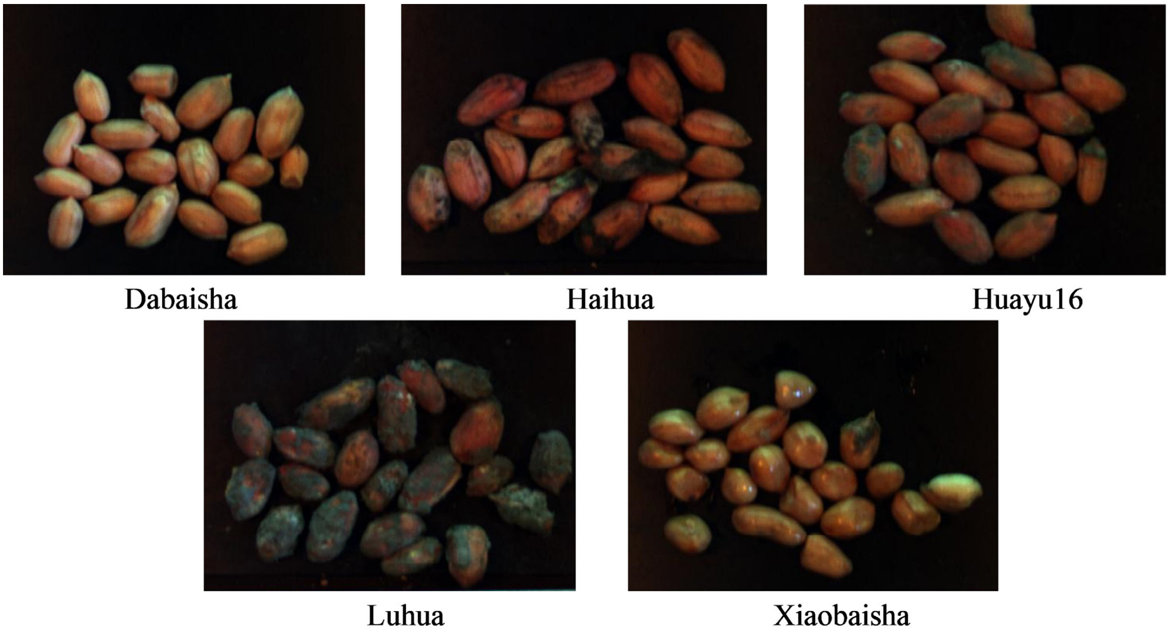

Figure 1. Moldy peanuts.

mold processing, a total of 200, put them in a closed box with cotton and a small amount of fresh water, and place them in an environment of 20 °C for 30 days. The remaining 400 peanuts, without any treatment, are also placed in a sterile and constant temperature condition of 20 °C for 30 days. After culture, spectral images were collected.

2.2 Hyperspectral image and data acquisition

The peanuts hyperspectral Images acquisition system adopts the image-λ "spectral Image" series hyperspectral machine of Zhuoli Hanguang Company, and SpacVIEW software is used to operate it. The system consists of a computer, a transmitter, a dot array camera and a halogen light source, as shown in Figure 2. The effective band range of its spectrum is 387-1034 nm, and the band resolution is 2.8 nm, with a total of 256 bands and 1344*1024 pixels. The determination and display attributes R, G and B of each group of samples were set as 638.7, 551.58 and 442.95, respectively, and the time was set as 10 s. The distance between the peanut sample and the camera lens was set as 165 mm, the sample movement speed was set as 4.7 mm·s-1, the exposure time of the camera was 4 ms, and the scanning area of the spectrum was 150 mm (Hong et al., 2015). In the process of collection, due to the influence of noise caused by the surrounding environmental factors and the dark current of the instrument, it is necessary to collect black and white frames respectively before sample collection, and conduct black-and-white correction in SpacVIEW according to the following formula (Equation 1) after sample collection (Yu et al., 2016). ENVI5.1 (The Environment for Visualizing Images) software was used to extract the areas required by the experiment in the peanut hyperspectral image after black and white correction, and then calculate the reflection average value of the spectral data in the extracted area as the characteristic reflection spectral curve of different peanut varieties.

$$R_1 = \frac{R_0 - R_B}{R_w - R_B} \tag{1}$$

Where, R_0 is the initial hyperspectral image (RAW), R_1 is the image after black-and-white correction, R_B is the black frame image collected after closing the lens, and R_w is the white frame image collected after correct collection and debugging without putting the sample.

2.3 Spectral data pre-processing

Because each ROI region of hyperspectral image extracted has 256 bands, corresponding to 256 data, the amount of data is large; And in hyperspectral data, weather, light, and the differences or artificial operation inevitable error, need to noise reduction of collected data, according to the strong correlations between the properties and spectral data band according to the amount of total high redundancy feature, so this research used a variety of data preprocessing algorithms, and compared their results, get the optimal modeling spectrum data, including: Min-Max Standardization, Boxing Smoothing, Wavelet Threshold Denoising, Exponential Smoothing, Median Filtering, Logarithmic Transformation Normalization, Z-Score Standardization, Local Regression-weighted linear least squares + polynomial model, Savitzky-Golay to eliminate noise in Raw spectral data.

Smoothing algorithm

Brown gives an exponential smoothing algorithm. It holds that the change of time order is regular and can reasonably continue the time series. Index smoothing includes primary, secondary and multiple indices smoothing (Huang, 2014).

Logarithmic transformation normalization

Some of the data changes are designed to be consistent with the assumptions we make so that we can analyze them in theory. Log transformation is a special method of data transformation, which transforms some common problems that we have not solved theoretically into solvable problems (Wang et al., 2012).

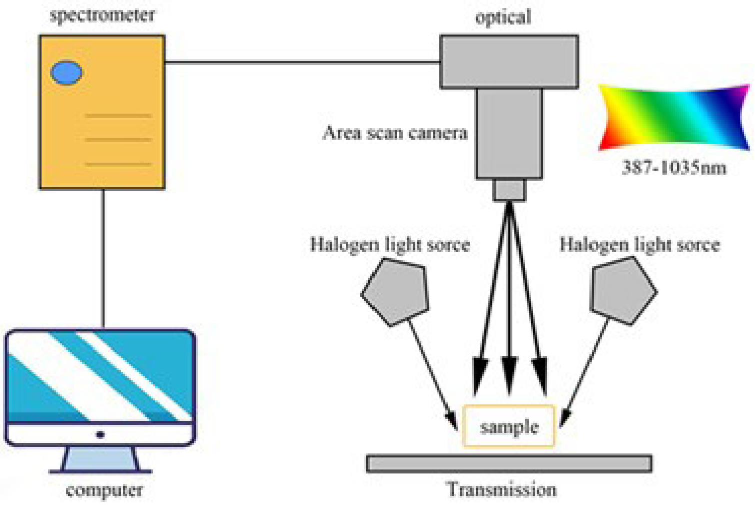

Figure 2. Image-λ "spectral Image" series high spectrometers of Zhuoli Hanguang Company.

Min-max standardization

0-1 standardization is mainly to change the data to the range of [0, 1] (Ji et al., 2016). Assuming that the sample is, the mathematical calculation formula of 0-1 standardization is as follows (Equation 2):

$$x_i^* = \frac{\max(x) - x_i}{\max(x) - \min(x)} \tag{2}$$

Wavelet threshold denoising

Information using wavelet transform method (using Mallat algorithm), as a result of the information generated in the wavelet coefficient is important information, so the wavelet coefficient information will increase after decomposition via wavelet transform coefficient, and the SNR value will decrease, and because the signal-to-noise ratio is less than the wavelet coefficients of information, by selecting an appropriate threshold, When the wavelet coefficient exceeds the threshold point, it will be regarded as the emergence and retention of information; when the wavelet coefficient approaches the threshold, it will be regarded as the noise caused by the removal, so as to remove the data causing errors in the experiment (Zhou, 2015). Its essence is to control the part that is not in the information or increase the useful information part.

The basic steps are as follows:

1. Decomposition: carry out wavelet analysis on the information based on the pre-determined wavelet with a certain layer number of N;
2. Threshold processing process: select appropriate threshold after analysis and measure the coefficient of each layer by threshold function;
3. Reconstruction: signal reconstruction with the processed coefficients.

Median filtering

Median Filtering is a nonlinear data preprocessing method. It was in the 1970s that J W Jukey invented and applied the data processing technology in one-dimensional information for the first time, and later it was widely applied in two-dimensional information processing. It is now widely used in image enhancement and data processing (Liu et al., 2019). Median Filtering can also be used in data pre-processing. Median filters generally assume that the data itself is stable and undistorted, and that all sudden changes caused by noise can be eliminated. The median filter also produces fuzzy signals and is most suitable for processing very independent and prominent noise information. Because the median filter is usually sliding and has an odd number of Windows. The specific steps are: Create a window of length 2N +1 and move the window bit by bit across the data sequence. After each move, the sequence of data information in the window is rearranged. Replace the small number of positions in the window before permutation with the medium number obtained after permutation. After arranging the values of the m points according to the number size, the number in the middle of the serial number is selected as the filter output value. Expressed by the following mathematical formula (Equation 3):

$$y = medfit1(x, n) \tag{3}$$

In this algorithm, y represents the reflectivity data corresponding to wavelength after one-dimensional median filtering, x represents the reflectivity data corresponding to original wavelength, and n represents the data width.

Z-score standardization

Zero-mean normalization is also called standard deviation normalization, which means that the average of the processed data is 0 and the standard deviation is 1 (Yang et al., 2022).

The transformation formula is (Equation 4):

$$x^* = \frac{x - \bar{x}}{\sigma} \tag{4}$$

Where $\bar{x}$ is the mean value of the original data, and σ is the standard deviation of the original data.

Local regression-weight linear least squares

Original linear regression objective function (Equation 5):

$$J(\theta) = \frac{1}{2}\Sigma_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \tag{5}$$

Locally weighted linear regression objective function (Equation 6):

$$J(\theta) = \frac{1}{2}\Sigma_{i=1}^{m}\omega^{(i)}(h_\theta(x^{(i)}) - y^{(i)})^2 \tag{6}$$

The difference between the two is that the latter has more weight $\omega^{(i)}$, which can control the influence of the prediction error of the *i-th* sample on the objective function. The idea of locally weighted linear regression is: when making prediction for a certain sample, focus on the nearby weights of the sample and give them higher $\omega^{(i)}$. Therefore, the specific form of $\omega^{(i)}$ is shown as follows (Zhang et al., 2013) (Equation 7):

$$\omega^{(i)} = \exp(-\frac{\left(x^i - x\right)^2}{2\tau^2}) \tag{7}$$

τ is the wavelength parameter. The larger τ is, the faster the distance sample falls, and the smaller the effect of distance sample.

Savitzky-Golay filtering

SG filter uses low order polynomial fitting of continuous subset of the same information point, and uses linear least square to predict the next time series data point. For example, if you specify a filter window consisting of five data points in the range 1 to 5, you can use linear least squares to derive n-order polynomial fits with those data points and use polynomials to infer a sixth data point. Next, the filter moves forward in time to cover the five data points in the range 2 to 6 to infer a seventh data point. This process continues until all data in the original set is filtered out (Lu et al., 2019).

The solution of Savitzky-Golay convolution smoothing matrix operator is an important step.

The width of the filtering window is $n = 2m + 1$ and each measuring point $x = (-m, -m+1, 0...0, 1, ...m-1, m)$ uses $k-1$ polynomial to fit the data points in the window (Equation 8):

$$y = a_0 + a_1 x + a_2 x^2 + ... + a_{k-1} x^{k-1} \tag{8}$$

So, you have n of these equations, which are subsumed into k linear equations. In order to make the equations have a solution, n should be greater than or equal to k, $n > k$ is generally selected, and the fitting parameter A is determined by least square fitting (Equation 9).

$$Y_{(2m+1)\times 1} = X_{(2m+1)\times k} A_{K\times 1} + E_{(2m+1)\times 1} \tag{9}$$

The least squares solution of A is (Equation 10):

$$A = (X^T X)^{-1} X^T Y \tag{10}$$

The model prediction or filtering value $\dot{Y}$ of Y is (Equation 11):

$$\dot{Y} = XA = X(X^T X)^{-1} X^T Y = BY \tag{11}$$

2.4 Feature extraction and modeling

Gradient Boosting Decision Tree (GBDT)

GDBT is the main method in our experiment. Gradient reinforcement: Gradient reinforcement is a machine learning technique that, when used for regression and analysis of problems, uses its weak estimation mode (generally decision tree) to obtain the result set in the form of estimation operator (Ni et al., 2009; Yu et al., 2021). Like other reinforcement algorithms, the model is built in the form of stages, and any separable loss function in the generalized model can be optimized (Equation 12).

$$f_0(x) = \arg\min_c \sum_{i=1}^{N} L_{(y_i, c)} \tag{12}$$

1. To $m = 1, 2, ..., M$

   a. For each sample $i = 1, 2, ..., N$, calculate the negative gradient as the residual (Equation 13).

$$\gamma_{im} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x) = f_{m-1}(x)} \tag{13}$$

   b. The residual obtained in the previous step is taken as the new true value of the sample, and the data (x_i, γ_{im}) and $i = 1, 2, ..., N$ are taken as the training data of the next tree to obtain the regression tree $f_m(x)$ and its corresponding leaf node region $R_{jm}, j = 1, 2, ..., J$。 Where J is the number of leaf nodes of regression tree t.

   c. Calculate the best fitting value for leaf region $j = 1, 2, ..., J$ (Equation 14).

$$\gamma_{jm} = \underbrace{\arg\min}_{r} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma) \tag{14}$$

   d. Update strong learner (Equation 15):

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}) \tag{15}$$

2. Get the final learner (Equation 16):

$$f(x) = f_M(x) = f_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}) \tag{16}$$

Extreme Gradient Boosting Decision Tree Algorithm (XGBoost)

XGBoost is an analysis and regression algorithm based on gradient boosting decision tree (GBDT). XGBoost firstly constructs a considerable number of weak learners, mainly classified regression trees, to train weak learners. It also completed the weighted calculation and summation after training to get the final regression model. During the construction process, start adding new educators based on the residual errors obtained in the last weak educator iteration. The new pedagogues are positioned on gradients to ensure that overall model deviation is reduced. Finally, a model with more important functions of regression and prediction is constructed. Compared with GBDT algorithm, XGBoost has many advances. XGBoost also introduces regularization terms for L1 and L2. In GBDT optimization modeling process, only the first derivative is used. XGBoost uses a second-order Taylor expansion for the loss function. XGBoost also supports column sampling to avoid computational workload reduction due to over-fitting. After each iteration, XGBoost allocates learning speed to leaf nodes, reducing the weight of each tree and providing better space for subsequent learning (Chen et al., 2016; Liu et al., 2021).

XGBoost's prediction model can be expressed as (Equation 17):

$$\widehat{y_l} = \sum_{K=1}^{K} f_k(x_i) \tag{17}$$

Where K is the total number of all trees, f_k is the *k-th* tree, and $\widehat{y_l}$ is the predicted result of sample x_i.

The XGBoost objective function is defined as (Equation 18):

$$Obj(\theta) \sum_{i=1}^{n} l(y_i, \widehat{y_l}) + \sum_{K=1}^{K} \Omega(f_k) \tag{18}$$

Where $l(y_i, \widehat{y_l})$ is the training error of sample x_i, and $\Omega(f_k)$ is the regular term of the *k-th* tree.

Two parts constitute the objective function. The first part measures the difference between predicted and true scores, and the other part is the regularization term. The regularization term also contains two parts, T represents the number of leaf nodes, w represents the score of leaf nodes. γ can control the number of leaf nodes, and λ can control the fraction of leaf nodes not to be too large to prevent over-fitting.

The newly generated tree is to fit the residual of the last prediction, that is, when t trees are generated, the predicted score can be written as (Equation 19):

$$\widehat{y_i}^{(t)} = \widehat{y_i}^{(t-1)} + f_t(x_i) \tag{19}$$

Meanwhile, the objective function can be written as (Equation 20):

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_{(t)}) \tag{20}$$

Obviously, the next step is to find a f_t that minimizes the target function. The basic idea of XGBoost is to approach it using its Taylor second order expansion at $f_t = 0$. Then, the objective function is approximately (Equation 21):

$$L^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \widehat{y_i}^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^{2}(x_i)] + \Omega(f_{(t)}) \tag{21}$$

Where g_i is the first derivative and h_i is the second derivative (Equation 22):

$$h_i = \partial^{\bar{2}}_{\widehat{y_i}^{(t-1)}} l(y_i, \widehat{y_i}^{(t-1)}) \tag{22}$$

Since the predicted scores of the first $t-1$ trees and the residual difference of y have no negative impact on the objective function, it can be removed. The simplified objective function is (Equation 23):

$$\tilde{L}^{(t)} \simeq \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^{2}(x_i)] + \Omega(f_{(t)}) \tag{23}$$

The above formula is to add up the loss function values of each sample, because we already know that all the data will eventually fall into a leaf node, so we should reorganize all the data blocks of the same leaf node. The steps are as follows (Equation 24):

$$\begin{aligned} Obj^{(t)} &\simeq \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^{2}(x_i)] + \Omega(f_{(t)}) \\ &= \sum_{i=1}^{n} [g_i \omega_q(x_i) + \frac{1}{2} h_i \omega_q^{2}(x_i)] + \gamma T + \lambda \frac{1}{2} \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) \omega_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) \omega_j^2] + \gamma T \end{aligned} \tag{24}$$

Therefore, by modifying the above formula, we can rewrite the objective function as a quadratic function about the leaf node fraction ω, and solve the optimal ω and objective function value becomes very simple, directly using the vertex formula can be.

Therefore, the optimal ω and objective function formula is (Equation 25):

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}, Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \lambda T \tag{25}$$

CategoricalBoost (CatBoost)

CatBoost is an algorithm based on gradient lifting. It has few optimization parameters and allows for faster training and testing. It uses the ordered boosting algorithm, which improves the generalization of the model.

CatBoost shows better results than random forest and other database-like gradient lifting algorithms. It is very suitable for categorizing data structures and avoids the need for intermediate data conversion (Al-Duwairi et al., 2020; Devi & Priya, 2021).

CatBoost has five main features:

1. High model accuracy can be obtained without adjusting parameters;
2. Classification variables can be supported without pre-processing of non-numerical class features;
3. GPU runs fast and expandable, which can be realized by gradient enhanced computing of one GPU during training model, supporting parallel computing;
4. In order to reduce overfitting, a new gradient lifting mechanism model is constructed;
5. Fast prediction speed.

Light Gradient Boosting Machine (LightGBM)

LightGBM (Ji et al., 2021; Zhao & Khushi, 2021) has three major optimizations on XGBoost:

1. Histogram algorithm: histogram algorithm;
2. GOSS algorithm: gradient unilateral sampling algorithm;
3. EFB algorithm: mutually exclusive feature binding algorithm.

The relationship between LightGBM and XGBoost can be shown using the following formula (Equation 26):

$$LightGBM = XGBoost + Histogram + GOSS + EFB \tag{26}$$

Due to the introduction of these three methods, the complexity required to produce leaf diagrams in LightGBM is greatly reduced and the computation time is greatly saved, while the calculation of histogram also converts functions from floating point to any integer from 0 to 255 bits for storage, thus greatly saving the internal reserve.

LightGBM is compared to XGBoost as follows:

LightGBM can also be considered an improved version of XGBoost, which is lighter to run on large data sets than XGBoost.

1. XGBoost's model accuracy is comparable to LightGBM;
2. LightGBM trains much faster than XGBoost;
3. LightGBM has much less memory consumption than XGBoost;

4. Both XGBoost and LightGBM can handle feature missing values automatically;
5. Classification features: XGBoost does not support category features, requiring OneHot coding preprocessing. LightGBM directly supports category features.

2.5 Model optimization

As a parameter tuning tool, optuna is suitable for most machine learning frameworks, sklearn, xgb, lgb, pytorch, etc. Using a record of suggested parameter values and evaluated objective values, the sampler basically keeps narrowing the search space until it finds an optimal search space that produces parameters that lead to better objective function values.

Optuna is a tool that automatically helps us debug parameters. Optuna is much easier to use than sklearn's gridsearchcv. One is that optuna can quickly adjust parameters compared to sklearn, and the other is that it can visualize the process of debugging parameters. At the same time, if you haven't finished training, you can continue training next time. And optuna uses the mechanism of Bayesian debugging parameters internally, which can give us a relatively good result in the shortest time, and may even get an optimal result (Akiba et al., 2019).

3 Results and discussion

3.1 Analysis of raw spectral data

After SpacVIEW black and white correction treatment, ENVI5.1 was used to intercept each peanut's region of interest and calculate its average reflection value. The spectral reflection value curve of the obtained data was drawn as shown in Figure 3. It can be seen that starting from the initial band of 387 nm, there was an absorption peak state, and the first wave peak appeared in the vicinity of 392 nm. Moldy Dabaisha's peak value is the maximum, health peanuts is the minimum, then reflectance declined precipitously, near 410 nm, 6 kinds of peanuts in decline after the first small peaks, near 440 nm, healthy peanut appeared a peak, 5 kinds of mildew of peanuts in the troughs, since spectrum curve showed a trend of rise, Xiaobaisha's rising trend is the smallest, while Dabaisha is the most high, from the above analysis, it can be seen that the difference between healthy peanuts and moldy peanuts in the initial band is large and easy to distinguish, indicating that it is feasible to detect moldy peanuts based on hyperspectral image technology.

3.2 Spectral data preprocessing

In order to eliminate the influence of non-quality factor information in hyperspectral spectral data, 10 spectral pretreatment methods were used to eliminate noise in the original spectral data, and each algorithm was evaluated, as shown in the table below. As can be seen from Table 1, the accuracy of LR1, SG and ZSS algorithms after processing is equal to the original data, which is 95.8%. The accuracy of WTD and MMS algorithms after processing is not as good as the original data and is initially eliminated. The accuracy of other algorithms is above the original data, among which MF algorithm is the highest, which is 97.7%. LR2 algorithm is the best in healthy peanuts, reaching 1. MF algorithm is the best in all 5 kinds of moldy-growing peanuts. In addition, MF algorithm is the best in Log Loss, Hamming Loss and Fit Time. Therefore, the pre-processed data are used for modeling detection. The five types of peanuts in the table below are all moldy peanuts.

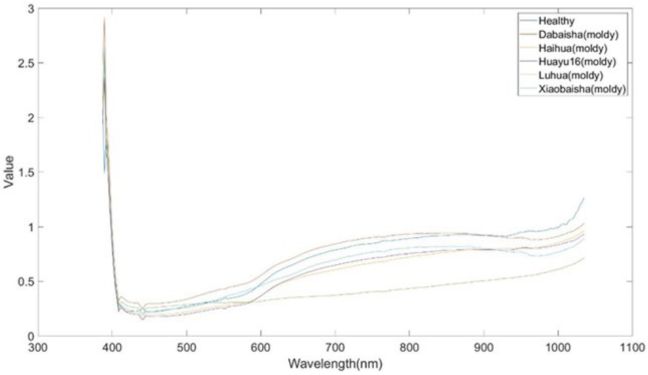

Figure 3. Spectral reflection curve.

Table 1. Spectral data preprocessing results.

Method	ACC (Accuracy)	Health	Dabaisha	Haihua	Huayu	Luhua	Xiaobaisha	Log Loss	Hamming Loss	Fit Time
MF	97.7%	0.9957	0.9310	0.8525	0.9459	0.9844	0.9000	6312	0.0226	0.88
LR2	96.1%	1.0000	0.8983	0.7241	0.8933	0.9147	0.8254	6796	0.0393	1.00
LTN	96.1%	0.9991	0.8983	0.7119	0.8919	0.9231	0.8387	6796	0.0393	0.92
BS	96.0%	0.9991	0.8908	0.7119	0.8933	0.9231	0.8197	6830	0.0405	0.96
ES	96.0%	0.9983	0.8814	0.7333	0.9007	0.9231	0.8197	6830	0.0405	0.99
LR1	95.8%	0.9983	0.8833	0.6897	0.9007	0.9231	0.8197	6865	0.0417	0.96
SG	95.8%	0.9991	0.8983	0.6780	0.8859	0.9147	0.8387	6865	0.0417	0.99
ORG	95.8%	0.9991	0.8983	0.6780	0.8859	0.9147	0.8387	6865	0.0417	1.01
ZSS	95.8%	0.9991	0.8983	0.6780	0.8859	0.9147	0.8387	6865	0.0417	1.01
WTD	95.7%	0.9991	0.9000	0.6667	0.8874	0.9063	0.8197	6899	0.0429	0.98
MMS	90.1%	0.9915	0.6812	0.5862	0.7286	0.8480	0.3404	8523	0.0988	1.17

3.3 Feature band extraction

GBDT, LightGBM, CatBoost and XGBoost algorithms are used to extract the top 30 feature bands. As shown in Figure 4, 1034.9899 nm is ranked in the top three in GBDT, LightGBM and CatBoost algorithms, indicating that among these three algorithms, In LightGBM and CatBoost algorithms, the top bands are the head-to-tail bands, which is consistent with the spectral curve analysis, indicating that in these two algorithms, the head-to-tail bands have better performance in mildew detection, while in XGBoost, the bands at 800 nm occupy the top five. From the extracted feature bands, it can be seen that the four algorithms have similarities and differences in mildew detection, indicating that each algorithm has its own special judgment for peanut mildew detection when it has common characteristics.

3.4 Modeling and optimization

Modeling and result analysis

In Table 2, GBDT, LightGBM, CatBoost and XGBoost algorithms were used to model the obtained feature bands. According to the results, the modeling accuracy of these features using GBDT is the lowest, 98%, and the operation time is the longest, all around 10 s. Among all the algorithms, the modeling accuracy of characteristic bands obtained by LightGBM is the highest, 99.1%, indicating that the first and last bands are the most effective for detection of peanut mildew. LightGBM takes the least time to model its characteristic bands, only 0.59 s, so in the model optimization, The LightGBM algorithm is selected for parameter tuning. The five types of peanuts in the table below are all moldy peanuts.

LightGBM optimization

Optuna algorithm was used to optimize LightGBM parameters (max_depth, n_estimator, num_leaves, subsample). The optimization process is shown in Figure 5. It can be seen from the isoline map that in the 300-time optimization process, the algorithm uses multi-fusion mode to select three parameters that are most suitable for a parameter. The four parameters are optimized in turn for 12 times, and the optimal parameter is obtained when the Objective Value reaches 1. The five types of peanuts in the table below are all moldy peanuts.

The optimized parameters are used for modeling, and the comparison with the original results is shown in the following Table 3:

As can be seen from the above table, from the perspective of accuracy, the optimization did not improve, but from the perspective of refinement, after optimization, the F-score of Haihua and Dabaisha decreased, and the Huayu and

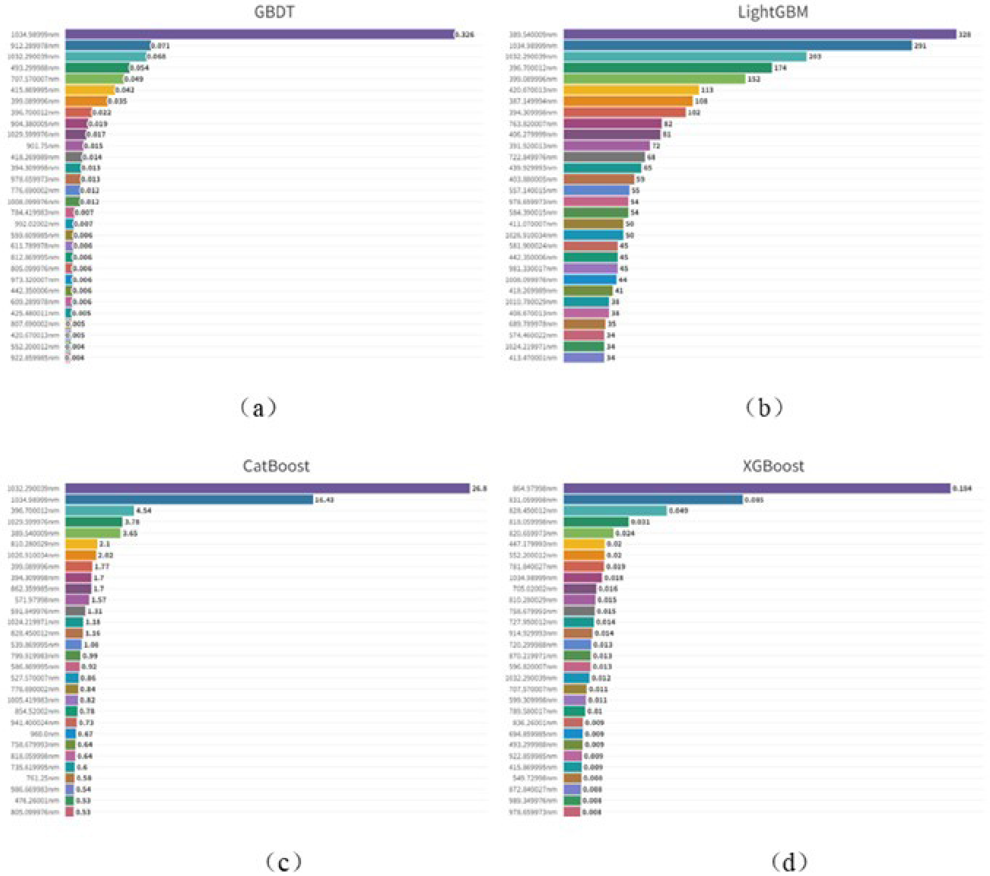

Figure 4. Extraction of important feature bands using (a) GBDT, (b) LightGBM, (c) CatBoost and (d) XGBoost.

Table 2. CatBoost, XGBoost, LightGBM and GBDT algorithms model performance metrics.

Model	Character	ACC	Health	Dabaisha	Haihua	Huayu	Luhua	Xiaobaisha	Log Loss	Hamming Loss	Fit Time
GBDT	**xgb**	98.9%	0.999	0.988	0.944	0.949	0.979	0.976	3751	0.011	0.98
	lgb	99.1%	1.000	1.000	0.941	0.951	0.989	0.950	3716	0.009	0.62
	cat	97.1%	0.996	0.953	0.750	0.911	0.957	0.900	4096	0.029	3.53
	gbdt	98.0%	0.997	0.976	0.872	0.917	0.979	0.927	3924	0.020	10.16
lgb	**xgb**	98.9%	0.999	0.988	0.944	0.949	0.979	0.976	3751	0.011	0.99
	lgb	99.1%	1.000	1.000	0.941	0.951	0.989	0.950	3716	0.009	0.59
	cat	97.1%	0.996	0.953	0.750	0.911	0.957	0.900	4096	0.029	3.67
	gbdt	98.0%	0.997	0.976	0.872	0.917	0.979	0.927	3924	0.020	10.20
cat	**xgb**	98.9%	0.999	0.988	0.944	0.949	0.979	0.976	3751	0.011	0.98
	lgb	99.1%	1.000	1.000	0.941	0.951	0.989	0.950	3716	0.009	0.64
	cat	97.1%	0.996	0.953	0.750	0.911	0.957	0.900	4096	0.029	3.54
	gbdt	98.0%	0.997	0.976	0.872	0.917	0.979	0.927	3924	0.020	10.17
xgb	**xgb**	98.9%	0.999	0.988	0.944	0.949	0.979	0.976	3751	0.011	0.96
	lgb	99.1%	1.000	1.000	0.941	0.951	0.989	0.950	3716	0.009	0.64
	cat	97.1%	0.996	0.953	0.750	0.911	0.957	0.900	4096	0.029	3.51
	gbdt	98.0%	0.997	0.976	0.872	0.917	0.979	0.927	3924	0.020	9.88

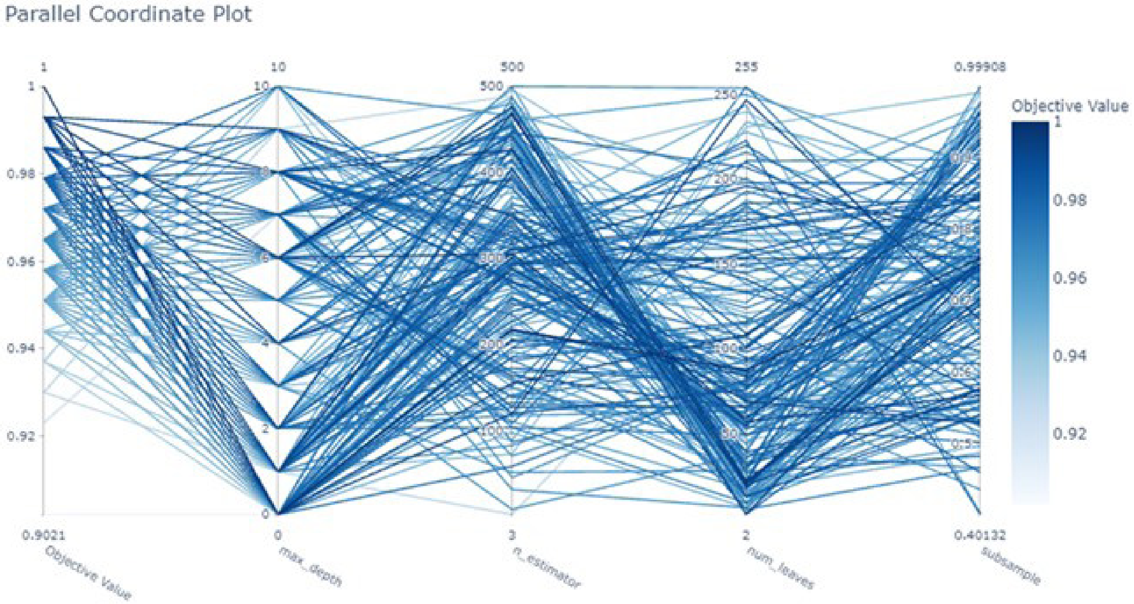

Figure 5. LightGBM optimization process.

Xiaobaisha increased. In the performance of fit time, the improvement was more obvious, with an increase of 42%. After optimization, the algorithm can improve the detection of mildew to a certain extent, especially in the running time, greatly reduce the time.

Result visualization

Modeling get confusion matrix using the optimized model as shown in Figure 6, in the confusion matrix can visually see that health has not been identified as the moldy peanuts, peanut moldy peanuts has not been mistakenly identified as health peanuts, prove model can accurately to mildew from health peanuts, but moldy Haihua have a was mistakenly identified the moldy Huayu, One was identified as moldy Luhua and another was identified as moldy Dabaisha; A moldy Huayu was mistakenly identified as a moldy Haihua; A small moldy Xiaobaisha was mistakenly identified as a moldy Huayu; There was no misidentification of moldy Luhua and Dabaisha.

Table 3. Optimization results.

Model	ACC	Health	Dabaish	Haihua	Huayu	Luhua	Xiaobaisha	Log Loss	Hamming Loss	Fitting Time
After	99.1%	1	0.988	0.895	0.970	0.989	0.977	3716	0.009	0.3413
Before	99.1%	1	1.000	0.941	0.951	0.989	0.950	3716	0.009	0.5899

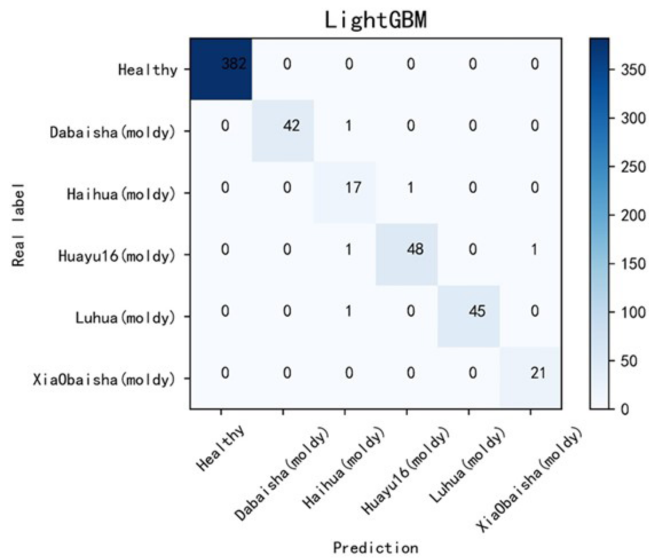

Figure 6. Confusion matrix after LightGBM optimization.

4 Conclusion

Hyperspectral imaging (387 nm-1034 nm) was used to study the mildew detection model of peanut during storage. Firstly, 120 peanuts of each type were selected, a total of 600 peanuts, and 40 peanuts of each type were randomly selected, and a total of 200 peanuts were treated with mildew. The remaining 400 peanuts were stored aseptically. After 30 days, the spectral images of peanuts were collected by Zhuoli Hanguang hyperspectral instrument, and the black-and-white correction was conducted by SpacVIEW. Then, the spectral image reflection data was extracted by ENVI5.1, and 10 pre-processing algorithms were used for denoising. Comprehensive comparison showed that Median Filtering (MF) had the best effect, and the recognition rate reached 97.7%. GBDT, LightGBM, CatBoost and XGBoost algorithms were used to extract the top 30 important feature bands in the spectral data after MF pretreatment. The feature bands extracted by the four algorithms were different from each other, but the top bands were all concentrated in the first and last bands, which proved that the first and last bands were of great significance to identify peanut mildew. GBDT, LightGBM, CatBoost and XGBoost algorithms were used to model the extracted feature bands. From the running results of the model, it can be seen that LightGBM is the optimal feature band extraction algorithm and model, and its detection rate can reach 99.1%. Optuna algorithm is used to tune its parameters. Compared with the traditional model, the operation error is greatly reduced and the operation time is also saved. Therefore, the optimal detection algorithm MF-LightGBM-LightGBM-Optuna-LightGBM was obtained, which provided theoretical and practical support for peanut mildew detection in the storage process.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: a next-generation hyperparameter optimization framework*. In A. Teredesai & V. Kumar (Orgs.), *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623-2631). New York: Association for Computing Machinery. http://dx.doi.org/10.1145/3292500.3330701.

Al-Duwairi, B., Al-Kahla, W., AlRefai, M. A., Abedalqader, Y., Rawash, A., & Fahmawi, R. (2020). SIEM-based detection and mitigation of IoT-botnet DDoS attacks. *Iranian Journal of Electrical and Computer Engineering*, 10(2), 2182-2191. http://dx.doi.org/10.11591/ijece.v10i2.pp2182-2191.

Borregaard, T., Nielsen, H., Nørgaard, L., & Have, H. (2000). Crop–weed discrimination by line imaging spectroscopy. *Journal of Agricultural Engineering Research*, 75(4), 389-400. http://dx.doi.org/10.1006/jaer.1999.0519.

Chen, T. C., & Yu, S. Y. (2022). Research on food safety sampling inspection system based on deep learning. *Food Science and Technology*, 42, e29121. http://dx.doi.org/10.1590/fst.29121.

Chen, T., Tong, H., & Benesty, M. (2016). *xgboost: extreme gradient boosting*. Vienna: The R Foundation.

Devi, R. R., & Priya, V. V. (2021). Multicollinear gradient catboost classification for enhance the preterm neonatal apnea level classification in medical data. *Materials Today: Proceedings*. In press.

Farah, J. S., Cavalcanti, R. N., Guimarães, J. T., Balthazar, C. F., Coimbra, P. T., Pimentel, T. C., Esmerino, E. A., Duarte, M. C. K. H., Freitas, M. Q., Granato, D., Cucinelli, R. P. No., Tavares, M. I. B., Calado, V., Silva, M. C., & Cruz, A. G. (2021). Differential scanning calorimetry coupled with machine learning technique: an effective approach to determine the milk authenticity. *Food Control*, 121, 107585. http://dx.doi.org/10.1016/j.foodcont.2020.107585.

Gonçalves, B. L., Uliana, R. D., Coppa, C. F. S. C., Lee, S. H. I., Kamimura, E. S., Oliveira, C. A. F., & Corassin, C. H. (2022). Aflatoxin M-1: biological decontamination methods in milk and cheese. *Food Science and Technology*, 42, e22920. http://dx.doi.org/10.1590/fst.22920.

Gonçalves, L., Rosa, A. D., Gonzales, S. L., Feltes, M. M. C., Badiale-Furlong, E., & Dors, G. C. (2017). Incidence of aflatoxin M-1 in fresh milk from small farms. *Food Science and Technology*, 37(Spe), 11-15. http://dx.doi.org/10.1590/1678-457x.06317.

Guo, Z. M., Guo, C., Wang, M. M., Shi, J. Y., Chen, Q. S., & Zou, X. B. (2019). Research advances in nondestructive detection of fruit and vegetable quality and safety by near infrared spectroscopy. *Shipin Anquan Zhiliang Jiance Xuebao*, 10(24), 8280-8288.

Hong, Y. W., Chen, Q., Zhang, J. S., & Ren, Y. P. (2015). Research progress of peanut allergens and its detection methods. *Shipin Anquan Zhiliang Jiance Xuebao*, 6(1), 226-233.

Hou, Y. C., Zhao, P. H., Zhang, F., Yang, S. R., 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, S. M. (2014). Design and implementation of automated test score platform based on B/S. *Computer Programming Skills & Maintenance*, (14), 12-13.

Idris, Y. M. A., Mariod, A. A., Elnour, I. A., & Mohamed, A. A. (2010). Determination of aflatoxin levels in Sudanese edible oils. *Food and Chemical Toxicology*, 48(8-9), 2539-2541. http://dx.doi.org/10.1016/j.fct.2010.05.021. PMid:20478351.

Ji, X. J., Du, S. B., & Wang, G. D. (2016). Using min-max normalization to measure the differences of regional economic growth—a case study of Yulin area,Shanxi province. *Economy and Management*, 30(3), 54-56.

Ji, X., Chang, W., Zhang, Y., Liu, H., Chen, B., Xiao, Y., & Zhou, S. (2021). Prediction model of hypertension complications based on GBDT and LightGBM. *Journal of Physics: Conference Series*, 1813(1), 012008. http://dx.doi.org/10.1088/1742-6596/1813/1/012008.

Kumar, P., Mahato, D. K., Kamle, M., Mohanta, T. K., & Kang, S. G. (2017). Aflatoxins: a global concern for food safety, human health and their management. *Frontiers in Microbiology*, 7, 2170. http://dx.doi.org/10.3389/fmicb.2016.02170. PMid:28144235.

Liu, Y., Wang, H., Fei, Y., Liu, Y., Shen, L., Zhuang, Z., & Zhang, X. (2021). Research on the prediction of green plum acidity based on improved XGBoost. *Sensors*, 21(3), 930. http://dx.doi.org/10.3390/s21030930. PMid:33573249.

Liu, Z. J., Xia, Y. H., Yang, D. Z., Lin, Y., & Chang-Bin, X. U. (2019). An improved method for infrared image noise processing based on median filter. *Laser Infrared*, 49, 376-380.

Liu, Z., Jiang, J., Qiao, X., Qi, X., Pan, Y., & Pan, X. (2020). Using convolution neural network and hyperspectral image to identify moldy peanut kernels. *LWT*, 132, 109815. http://dx.doi.org/10.1016/j.lwt.2020.109815.

Lu, Y.-B., Liu, W.-Q., Zhang, Y.-J., Zhang, K., He, Y., You, K., Li, X.-Y., Liu, G.-H., Tang, Q.-X., Fan, B.-Q., Yu, D.-Q., & Li, M.-Q. (2019). An adaptive hierarchical Savitzky-Golay spectral filtering algorithm and its application. *Guang Pu Xue Yu Guang Pu Fen Xi*, 39(9), 2657-2663.

Mutua, F., Lindahl, J., & Grace, D. (2019). Availability and use of mycotoxin binders in selected urban and peri-urban areas of Kenya. *Food Security*, 11(2), 359-369. http://dx.doi.org/10.1007/s12571-019-00911-4.

Nakariyakul, S., & Casasent, D. P. (2011). Classification of internally damaged almond nuts using hyperspectral imagery. *Journal of Food Engineering*, 103(1), 62-67. http://dx.doi.org/10.1016/j.jfoodeng.2010.09.020.

Ni, S., Chen, D., & Lv, M. (2009). *Research on optimization model of initial schedule of passenger trains based on improved genetic algorithm*. In B. Xie (Org.), *Second International Conference on Intelligent Computation Technology & Automation* (pp. 273-276). New York: Institute of Electrical and Electronics Engineers.

Pranoto, W. J., Al-Shawi, S. G., Chetthamrongchai, P., Chen, T. C., Petukhova, E., Nikolaeva, N., Abdelbasset, W. K., Yushchenko, N. A., & Aravindhan, S. (2022). Employing artificial neural networks and fluorescence spectrum for food vegetable oils identification. *Food Science and Technology*, 42, e80921. http://dx.doi.org/10.1590/fst.80921.

Sun, X. D., Su, P., & Shan, H. (2017). Mycotoxin contamination of rice in China. *Journal of Food Science*, 82(3), 573-584. http://dx.doi.org/10.1111/1750-3841.13631. PMid:28135406.

Wang, Q. J., Shrestha, D. L., Robertson, D. E., & Pokhrel, P. (2012). A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research*, 48(5), W05514. http://dx.doi.org/10.1029/2011WR010973.

Yang, Y., Liu, S. X., Xing, B., & Li, K. S. (2022). Face inpainting via Learnable Structure Knowledge of Fusion Network. *Ksii Transactions on Internet and Information Systems*, 16(3), 877-893.

Yang, J. D., Hainaut, P., Gores, G. J., Amadou, A., Plymoth, A., & Roberts, L. R. (2019). A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nature Reviews. Gastroenterology & Hepatology*, 16(10), 589-604. http://dx.doi.org/10.1038/s41575-019-0186-y. PMid:31439937.

Yu, H.-W., Wang, Q., Liu, L., Shi, A.-M., Hu, H., & Liu, H.-Z. (2016). Research process on hyperspectral imaging detection technology for the quality and safety of grain and oils. *Guang Pu Xue Yu Guang Pu Fen Xi*, 36(11), 3643-3650. PMid:30199206.

Yu, Z., Wang, Z., Zeng, F., Song, P., Baffour, B. A., Wang, P., Wang, W., & Li, L. (2021). Volcanic lithology identification based on parameter-optimized GBDT algorithm: a case study in the Jilin Oilfield, Songliao Basin, NE China. *Journal of Applied Geophysics*, 194, 104443. http://dx.doi.org/10.1016/j.jappgeo.2021.104443.

Yuan, D., Jiang, J., Qi, X., Xie, Z., & Zhang, G. (2020). Selecting key wavelengths of hyperspectral imagine for nondestructive classification of moldy peanuts using ensemble classifier. *Infrared Physics & Technology*, 111, 103518. http://dx.doi.org/10.1016/j.infrared.2020.103518.

Zhang, H. Z., Luo, H. M., & Luo, H. (2013). Weighted least square solution to generalized anti-skew-symmetric matrices on the linear manifold. *Journal of Xihua University*, 2, 25-28.

Zhang, Y.-Z., Xu, M.-M., Wang, X.-H., & Wang, K.-Q. (2019). Hyperspectral image classification based on hierarchical fusion of residual networks. *Guang Pu Xue Yu Guang Pu Fen Xi*, 39(11), 3501-3507.

Zhao, Y., & Khushi, M. (2021). Wavelet denoised-ResNet CNN and LightGBM method to predict forex rate of change. *arXiv*. Ahead of print.

Zhou, Q., Acharya, G., Zhang, S., Wang, Q., Shen, H., & Li, X. (2016). A new perspective on universal preconception care in China. *Acta Obstetricia et Gynecologica Scandinavica*, 95(4), 377-381.

Zhou, Y. (2015). *Research on the applications of data mining in financial prediction*. In J. Wang & Y. Qin (Eds.), *2015 International Conference on Education Technology, Management and Humanities Science* (pp. 448-453). Dordrecht: Atlantis Press. http://dx.doi.org/10.2991/etmhs-15.2015.102.