ACTA SCIENTIARUM POLONORUM
Acta Sci. Pol. Technol. Aliment. 22(2) 2023, 169–181
www.food.actapol.net pISSN 1644-0730 eISSN 1898-9594 http://dx.doi.org/10.17306/J.AFS.2023.1124

ORIGINAL PAPER

Received: 8.03.2023
Accepted: 28.04.2023

PHASE DIAGRAM COMBINED WITH IMPROVED FUZZY SUPPORT VECTOR MACHINE FOR RAPID AND NONDESTRUCTIVE DETECTION OF DIARRHETIC SHELLFISH POISONING

Wei Jiang✉, Yao Liu, Fu Qiao, Zhongyan Liu, Jianfang Xiong, Shaogeng Zeng

Lingnan Normal University
Zhanjiang 524048, **China**

ABSTRACT

Background. The diarrhoeal shellfish poisoning (DSP) toxin is a powerful marine biological toxin. Eating DSP toxin-contaminated mussels will lead to serious gastrointestinal diseases. To this end, a method for the detection of DSP toxins using near-infrared reflectance spectroscopy combined with pattern recognition is proposed.

Material and methods. In the range from 950–1700 nm, spectral data of healthy mussels and DSP-contaminated mussels were obtained. To select the optimal band subsets, a band selection algorithm based on model cluster analysis was applied. As distinguishing DSP toxin-contaminated mussels from healthy mussels is a classification problem of imbalanced data, an improved fuzzy support vector machine-based recognition method was proposed. The influence of the parameters of the band selection algorithm and the fuzzy support vector machine on the model performance was analyzed.

Results. Compared with the traditional support vector machine, the proposed model has better performance in detecting DSP toxins and is not affected by the imbalance ratio. Its geometric mean value can reach 0.9886 and the detection accuracy can reach 98.83%.

Conclusion. The results show that as an innovative, fast and convenient analytical method, near-infrared spectroscopy is feasible for the detection of DSP toxins in mussels.

Keywords: near infrared spectroscopy, diarrheal shellfish, mussel, band selection, imbalanced classification

INTRODUCTION

Diarrhetic shellfish poisoning (DSP) is one of the marine biological toxins with the highest frequency and the widest distribution in coastal waters around the globe (Blanco et al., 2007). Okadaic acid (OA) is the most important component of DSP, and human consumption of mussels contaminated with high concentrations of OA toxins will cause diarrhetic poisoning, or even death in severe cases. The main symptoms include gastrointestinal disorders, diarrhea, abdominal pain, fever, vomiting and nausea (Corriere et al., 2021; Mak et al., 2005). It has been shown that eating DSP-contaminated mussels may increase the risk of digestive system cancer (Manerio et al., 2008). It should be noted that the nature of DSP toxins will not change for heating or freezing, and the taste of DSP-contaminated mussels will also not be affected. If there is no professional detection instrument, it is usually difficult to find toxin-contaminated mussels. As a result, extensive attention has been given to how to detect DSP toxin-contaminated mussels effectively and quickly.

✉jiangliehui@yeah.net, phone 86 0759 3183189

Traditional DSP toxin detection methods include mouse bioassay, enzyme-linked immunosorbent assay (Du et al., 2020), protein phosphatase inhibition assay (Chen et al., 2018) and high-performance liquid chromatography tandem mass spectrometry (Ioutsi et al., 2022). Although these methods are accurate, they are not suitable for daily monitoring because of their complicated operation, high cost, destructiveness, and lack of fast detection capability. Consequently, it is necessary to develop a simple, real-time, efficient and low-cost DSP toxin detection technology.

In view of the shortcomings and limitations of these methods, near-infrared (NIR) spectroscopy combined with chemometrics and machine learning methods has been explored as an alternative method for DSP toxin detection. It can provide fast, accurate, simple and reliable food quality and safety measurement methods, and has been used to detect meat, fruits and vegetables, seafood and rice (Dirks and Poole, 2022; Melado-Herreros et al., 2022; Savoia et al., 2020; Srivastava and Mishra, 2022). NIR spectroscopy has been successfully applied for the qualitative identification and quantitative analysis of shellfish, which can quickly estimate the moisture and glycogen content in eastern oysters (*C. virginica*; Guévélou and Allen, 2016), classify healthy *Tegillarca granosa* and *Tegillarca granosa* contaminated with heavy metals (Cu, Cd, Pb, Zn; Chen et al., 2015), quantitatively detect bivalve protein, lipid and glycogen composition (Bartlett et al., 2018), estimate the heritability of meat composition traits of Pacific oyster gold-shell strain(Wan et al., 2020), quantitatively detect marine Parkinworm infection levels in eastern oysters (Guévélou et al., 2021), and rapidly detect the mussels contaminated with Cd, Zn, Pb and Cu (Liu et al., 2022a; Xiong et al., 2022). There are few studies using NIR spectroscopy to detect DSP toxins contaminated mussels. Liu et al. (2022b) initially used NIR spectroscopy to detect DSP-contaminated mussels, and the results showed the feasibility of NIR spectroscopy to detect DSP toxins in mussels. Although the multilayer perceptron model based on first-derivative spectrum preprocessing obtains an optimal classification effect, when the imbalance ratio of unbalanced dataset is too large, the classification accuracy is not ideal. The application of full-spectrum modeling not only involves costly and too large computation, but also affects the stability of the model. The aim of this study is to evaluate the feasibility of the rapid detection of DSP toxins by NIR spectroscopy while minimizing the handling of mussel samples.

NIR spectral data contains a lot of information of a sample, and in some cases, part of it is irrelevant or redundant. When the learning model faces a large number of input features, the prediction performance will decline. There it is necessary to select the characteristic wavelength most relevant to the subject of the research. To obtain optimal prediction accuracy, the wavelength selection algorithm selects as few wavelengths as possible, because the less data there is to be analyzed, the faster the model is learned. Nowadays, many wavelength selection algorithms have been proposed (Raghavendra et al., 2021). Here a new calculation method – the Phase diagram (PHADIA) − is used to compress the spectral data of shellfish toxin samples, which can rapidly detect mussels contaminated with DSP toxins.

In practical applications of DSP toxin detection, the number of DSP toxins contaminated mussels is lower than that of healthy mussels. Traditional classification models usually assume that the training samples are evenly distributed in inter classes. Training the models thus requires a sufficient number of healthy mussel samples and corresponding DSP toxin-contaminated mussel samples. However, it is difficult to obtain enough contaminated samples to calibrate the detection model in practice, and the number of healthy and contaminated mussels is always unbalanced in the real environment. Without considering the imbalance problem, the classification algorithm favors most classes, i.e. the classifier may classify most of the samples as healthy samples. As a result, it is important to find a solution for unbalanced datasets. Fuzzy support vector machine (FSVM; Lin and Wang, 2002) is a classification method based on traditional support vector machine (SVM). It has shown optimal processing ability on unbalanced data, but the classification effect is obviously different under different imbalance ratios. To this end, an improved fuzzy support vector machine (IFSVM) algorithm is introduced to distinguish healthy mussels and DSP-contaminated mussels, aiming to solve the classification problem of sample imbalance in rapid nondestructive detection of shellfish toxins.

MATERIALS AND METHODS

Preparation of mussel samples

The mussels were purchased from Dongfeng Seafood Market, Zhanjiang, China. Mussels of similar sizes were selected and domesticated in plastic containers to adapt to the experimental environment. After 3 days of acclimation, mussels with higher vitality were selected for subsequent tests. The mussels that were selected were then transferred into two 119 cm × 108 cm × 32 cm plastic containers. Each water tank was filled with 80 L seawater, with a salinity of 30‰ and a temperature of 26°C. The daily feeding concentration of mussels in the experimental group was 7.3×10^9 cell/L of *Proorocentrum lima*, and the mussels in the control group were fed with 10^9 cell/L of photosynthetic bacteria every day. The feeding amount of *Proorocentrum lima* and photosynthetic bacteria was determined through preliminary experiments to maintain the healthy state of mussels.

In the experiment, the seawater in the breeding pool was continuously inflated to keep the mussels in optimal physiological condition. The seawater was replaced every 24 hours to keep the living environment of mussels. The experiment lasted for 6 days to accumulate DSP toxins in mussel samples, and a total of 240 samples (120 samples in each group) were collected for spectrum collection.

Obtaining spectra of mussel samples

The NIR spectroscopy of each mussel sample was obtained by using the NIR spectroscopy measurement system (Fig. 1) built by the laboratory. The system consists of a NIR spectrometer, a halogen light source, a Y-shaped optical fiber, a USB data line, an adjustable displacement platform and a computer. The NIR spectrometer model is SW2520-050-NIRA, produced by OtO Optoelectronics Co., Ltd. in Taiwan, China. The NIR spectroscopy includes 114 wavebands, ranging from 950 nm to 1700 nm, with an interval of 6.5 nm. Before collecting the NIR spectroscopy of the mussels, black-and-white correction was conducted to reduce noise (Liu et al., 2022c).

Taking the mussels from the seawater container, the fiber optic probe was positioned directly on the center surface of the mussels for spectral measurements. The spectrum of each sample is the average of three scans in reflection mode. The spectral data of the sample is acquired via SpectraSmart software.

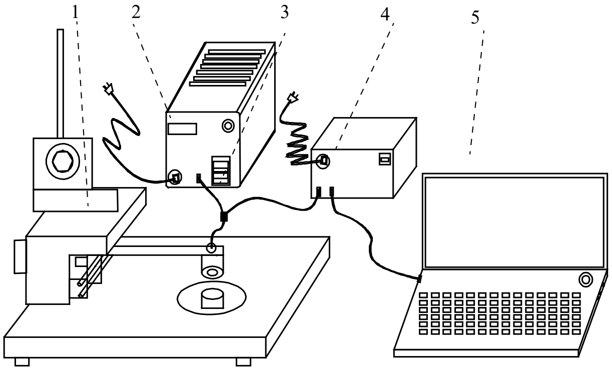

Fig. 1. Schematic diagram of NIR spectroscopy system: 1 – adjustable shifting platform, 2 – halogen light source, 3 – *Y*-shaped optical fiber, 4 – spectrograph, 5 – computer

After spectrum collection, samples of DSP-contaminated mussels in the experimental group were detected for DSP content using the LC-MS/MS method, and the detection results showed that the DSP content was 35 μg/kg.

Band selection algorithm

There is a large amount of interference and invalid information in the near-infrared spectral data, and it is necessary to select the most relevant information for DSP toxin detection. In this study, the PHADIA band selection algorithm based on model cluster analysis (Li et al., 2017) was applied to find the optimal subset. The flowchart of PHADIA band selection algorithm based on model cluster analysis is shown in Figure 2.

The PHADIA algorithm uses two metrics to evaluate the predictive performance of variables and project them into four regions in a two-dimensional plot, named phase plots. The main steps are as follows:

Given a spectral dataset (X, y), let X be a matrix of $n \times p$, the rows contain n samples, and the columns contain p-dimensional vectors. y is a vector of $n \times 1$, representing the classification label of each sample, with a value of 1 or −1.

1. Sub-dataset sampling of variable space. At each sampling, Q in the p variables is randomly sampled to produce a sub-dataset of $n \times Q$. This process is repeated N times to obtain N sub-datasets, denoted as (X_{sub}, y), $i = 1, 2, \ldots, N$. The choice of

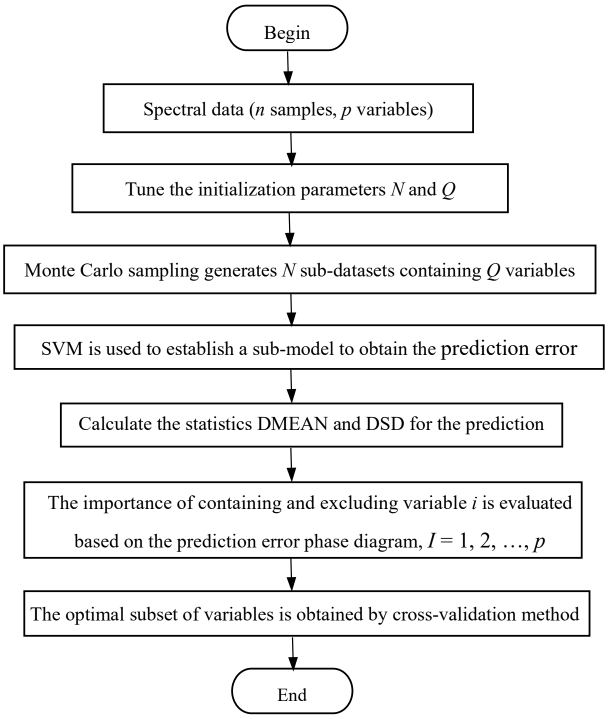
Begin
Spectral data (*n* samples, *p* variables)
Tune the initialization parameters *N* and *Q*
Monte Carlo sampling generates *N* sub-datasets containing *Q* variables
SVM is used to establish a sub-model to obtain the prediction error
Calculate the statistics DMEAN and DSD for the prediction
The importance of containing and excluding variable *i* is evaluated based on the prediction error phase diagram, *I* = 1, 2, …, *p*
The optimal subset of variables is obtained by cross-validation method
End

Fig. 2. The flowchart based on the PHADIA algorithm

Q value depends on several factors, such as sample size and computational cost. Here Q is optimized by cross-validation.

2. Establishment of SVM model. SVM classifier is used to build N sub-models, and the corresponding prediction error is recorded. The performance of each model is evaluated with 5-fold cross-validation.
3. Statistical analysis of calculating phase diagram prediction error. In order not to lose the generality, the i-th variable is used as an example to illustrate the process of calculating the phase diagram. The N SVM sub-models in (2) are divided into groups A (the model containing the i-th variable) and B (the model containing the i-th wavelength variable). Based on these two sets of prediction error data, the corresponding mean and standard deviation are calculated, which are expressed as MEAN_A, SD_A, MEAN_B and SD_B, respectively. Finally, the predictive power of the i-th variable is evaluated by calculating two statistics.

The first statistic is defined as the difference between $\text{MEAN}_{i,A}$ and $\text{MEAN}_{i,B}$, which is calculated as follows:

$$\text{DMEAN}_i = \text{MEAN}_{i,B} - \text{MEAN}_{i,A} \tag{1}$$

If $\text{DMEAN}_i > 0$, it indicates that the model containing the i-th variable has better predictive performance and vice versa. To determine the difference between the mean error values of group A and group B models, p is calculated using the nonparametric Mann-Whitney U test, combined with DMEAN, to determine whether the variables can significantly improve the prediction performance.

Another statistic is defined as:

$$\text{DSD}_i = \text{SD}_{i,B} - \text{SD}_{i,A} \tag{2}$$

If $\text{DSD}_i > 0$, it indicates that the model containing the i-th variable has better stability.

For each variable, based on the values of DMEAN and DSD, a two-dimensional plan diagram, named a phase diagram, is drawn, which visually displays the predicted performance of all variables in one view. The phase diagram is shown in Figure 3, and all variables are projected into four regions. The peak "1" of each region indicates the prediction error distribution of the model containing the variable, and the peak "0" indicates the prediction error distribution of the model that does not contain the variable. In phase 1, DMEAN > 0 and DSD > 0 can not only improve the performance of

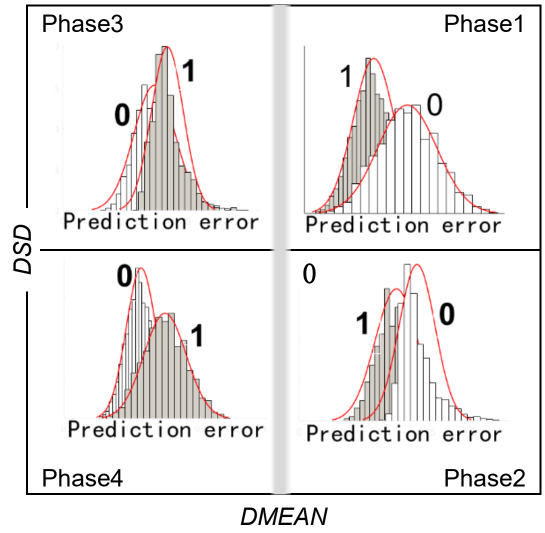

Fig. 3. Phase diagram of variable selection

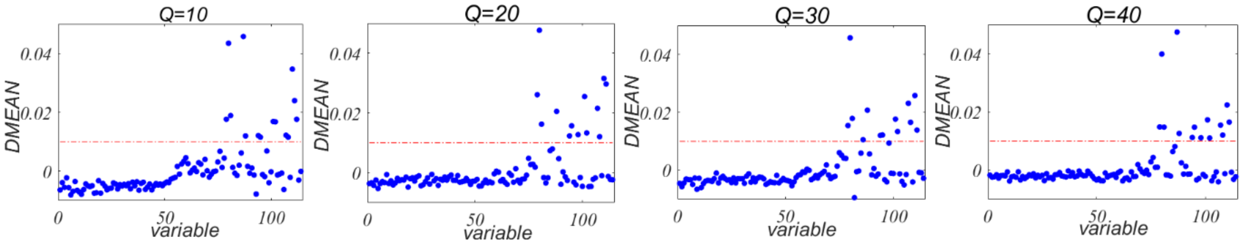

Fig. 4. DMEAN values of mixed datasets of DSP-contaminated samples and healthy samples at different Q values

the model (DMEAN > 0), but also reduce the prediction deviation (DSD > 0), which is called the information variable. In phase 2, DMEAN > 0 and DSD < 0 can improve the performance, but at the cost of increasing variance, which is not as good as the variable in phase 1. In phase 3, DMEAN < 0 and DSD > 0 decreases the performance of the model and reduces the variables of the prediction. In phase 4, DMEAN < 0 and DSD < 0 not only reduces the performance of the model, but also increases the prediction error.

To establish whether the PHADIA method could effectively filter the characteristic waves, the DMEAN of each band with the value Q is drawn, as shown in Figure 4. It can be seen that in all the Q values, the DMEAN value of useful information variables is the highest and obviously greater than zero. Conversely, the DMEAN values of the noise and interference variables are randomly concentrated around zero (the red dotted line in Figure 4 separates the informational wavelength variable from the other wavelength variable). It shows that the PHADIA method can screen out the wavelength variables that distinguish DSP contamination from healthy samples, and is robust with regard to Q values.

IFSVM classification model

In many machine learning methods, SVM has obvious advantages in complex nonlinear and high-dimensional spatial classification. However, the classification effect of traditional SVM on unbalanced datasets is not ideal. The SVM algorithm is biased towards the classification accuracy of most classes, while the classification effect is often poor on minority classes. In recent years, Lin and Wang (2002) proposed an FSVM method that applies fuzzy mathematics to SVM to overcome the influence of noise on support vectors to improve the accuracy of classification.

FSVM algorithm. For two classification problems, the supposition is the training set $(X,Y) = \{(x_i,y_i), i = 1, 2, \ldots, n\}$, where x_i is the sample, and y_i is the class label of x_i, $y_i \in \{-1, 1\}$. It is assumed that the first k samples are positive samples (i.e. $y_i = 1, i = 1, 2, \ldots, k$), and the remaining $n - k$ samples are negative samples (i.e. . $y_i = -1, i = k + 1, k + 2, \ldots, n$).

The general form of FSVM for unbalanced data classification is expressed as:

$$\min_{\omega,\xi} \tfrac{1}{2}\|\omega\|^2 + C^+\sum\nolimits_{i=1}^{k} s_i^+\xi_i^2 + C^-\sum\nolimits_{i=k+1}^{n} s_i^-\xi_i^2 \qquad (3)$$

$$s.t.\ y_i(\omega^T\Phi(x_i) + b) \ge 1 - \xi_i,\ \xi_i \ge 0,\ i = 1, 2, \ldots, n$$

where $\Phi(x_i)$ is the nonlinear mapping; $\xi_i(i = 1, 2, \ldots, n)$ is the relaxation variable; C^+ and C^- are the penalty factors of positive and negative samples, respectively, indicating the imbalance between the two classes; s_i^+ and s_i^- are the membership functions of positive and negative samples respectively, indicating the importance of the sample in its class.

Design of FSVM fuzzy membership function. Lin and Wang (2002) defined the fuzzy membership function as:

$$s_i^+ = 1 - \frac{d_i^{cen^+}}{\max\limits_j\left(d_j^{cen^+}\right) + \delta},\ i = 1, 2, \ldots, k \qquad (4)$$

$$s_i^- = 1 - \frac{d_i^{cen^-}}{\max\limits_j\left(d_j^{cen^-}\right) + \delta},\ i = k + 1, k + 2, \ldots, n \qquad (5)$$

where $d_i^{cen^+} = \left\|x_i - \frac{1}{k}\sum\nolimits_{j=1}^{k} x_j\right\|$, $d_i^{cen^-} = \left\|x_i - \frac{1}{n-k}\sum\nolimits_{j=p+1}^{n} x_j\right\|$. δ is a small positive number to ensure that the fuzzy membership is greater than 0. However, when the distribution of the dataset is uneven, the method is likely to train the noise as normal positive and negative classification samples, resulting in a decrease in the overall classification accuracy of the algorithm.

To effectively reduce the impact of noise contained in the sample set and better solve the problem of unbalanced classification of the dataset, when designing the fuzzy membership function, the penalty factor is considered in detail, as are the influence of the distance between the training sample and its class center, the tightness around the sample, and the amount of information of the sample.

IFSVM. According to the *K*-nearest neighbor criterion, the compactness around the sample is defined as follows:

$$D_i^{+(-)} = \tfrac{1}{k}\sum\nolimits_{x_j \in N_K^{+(-)}(x_i)} \left\| x_i - x_j \right\|,\ i = 1, 2, \ldots, n \qquad (6)$$

$N_K{}^{+(-)}(x_i)$ is the set of *K*-near neighbors of the *i*-th sample of the positive (negative) class. Obviously, the smaller the value of $D_i{}^{+(-)}$ of a sample, the greater the possibility that the sample belongs to the positive (negative) class.

The samples with a large amount of information are given a large membership function value. The evaluation method of the sample information is shown as follows:

$$\varphi\left(x_i^{+(-)}\right) = -\left\| \omega^* \cdot x_i^{+(-)} + b \right\|,\ i = 1, 2, \ldots, k \ (7)$$

where ω^* and b represent the normal vector and threshold of the classification hyperplane of traditional SVM, respectively; $\varphi(x_i{}^{+(-)})$ is the amount of information of the *i*-th positive (negative) sample. It is clear that the smaller the amount of information in the positive sample $\varphi(x_i{}^{+})$, the larger the corresponding sample information; on the contrary, the larger the negative class sample $\varphi(x_i{}^{-})$, the larger the corresponding amount ofinformation.

The membership of IFSVM is defined as follows:

$$s_i^+ = \frac{\max\left(\varphi_i^+\right) - \min\left(\varphi_i^+\right)}{\dfrac{q}{\varphi_i^+ - \min\left(\varphi_i^+\right) + 1} + \delta} * \left(1 - \alpha \frac{d_i^{cen^+}}{\max\limits_j\left(d_j^{cen^+}\right) + \delta} - (1-\alpha)\frac{D_i^+ - \min\left(D_i^+\right)}{\max\left(D_i^+\right) - \min\left(D_i^+\right) + \delta} \right) \qquad (8)$$

$$i = 1, 2, \ldots, k$$

$$s_i^- = E * \frac{\max\left(\varphi_i^-\right) - \min\left(\varphi_i^-\right)}{\varphi_i^- - \min\left(\varphi_i^-\right) + \delta} * \left(1 - \alpha \frac{d_i^{cen^-}}{\max\limits_j\left(d_j^{cen^-}\right) + \delta} - (1-\alpha)\frac{D_i^- - \min\left(D_i^-\right)}{\max\left(D_i^-\right) - \min\left(D_i^-\right) + \delta} \right)^M \qquad (9)$$

$$i = k + 1, k + 2, \ldots, n$$

where $\alpha \in [0,1]$ is the weight which is used to balance the importance of the near neighborhood density of the sample to the class center and the sample; the meaning of δ is similar to δ in Eqs. 4 and 5; M is used to adjust the range of fuzzy membership function of all samples, $M \in (0,1]$. The smaller the amount of information $\varphi_i{}^{+}$ of the *i*-th positive sample, the larger the corresponding sample information. The larger the amount of information $\varphi_i{}^{-}$ of the *i*-th negative sample, the larger the corresponding information. E is a balance factor that is used to ensure that the range of positive and negative class membership values is consistent. It is defined as the mean of the influence value of all training samples of the positive class divided by the mean of the influence value of all training sample information of the negative class. The formula is expressed thus (10):

$$E = \left(\sum\nolimits_{i=1}^{k} \frac{\max\left(\varphi_i^+\right) - \min\left(\varphi_i^+\right)}{\dfrac{1}{\varphi_i^+ - \min\left(\varphi_i^+\right) + 1} + \delta} * (n-k) \right) \Bigg/ \left(\sum\nolimits_{i=k+1}^{n} \frac{\max\left(\varphi_i^-\right) - \min\left(\varphi_i^-\right)}{\varphi_i^- - \min\left(\varphi_i^-\right) + \delta} * k \right) \qquad (10)$$

Parameter optimization. To make better use of the IFSVM algorithm to classify the actual data, it is necessary to optimize and select the parameters, such as δ, α, M, K, and C. In the experiment, the Radial basis function (RBF) kernel function is used, and the parameter γ need to be optimized and selected. Here the initial value of δ is selected by multiple experiments. δ is 0.0001; the range of α and M are {0, 0.1, …, 1} and {0.1, 0.2, …, 1}, respectively. Considering the training time of the algorithm, K is set to 6. According to the experimental results obtained by Ganaie and Tanveer (2021), when C^+/C^- is the ratio $((n - k) / k)$ of the number of the sample of the majority class to the minority class, the SVM algorithm can obtain better classification results. C^+ is set to $C(n - k)/k$, and C^- is $C(C > 0)$. The grid search method is used to select the penalty factor C and γ core parameters, and the ranges of C and γ are $\{2^{-1}, 2^0, 2^1, \ldots, 2^{10}\}$ and $\{2^{-15}, 2^{-13}, \ldots, 2^{-1}, 2^0\}$, respectively.

Evaluation method of unbalanced dataset classification

To overcome the disadvantages of single classification accuracy, the geometric mean (*Gmean*) is introduced to reflect the performance of the classifier on unbalanced datasets. *Gmean* is used to characterize the classifier's ability to correctly identify all classes of samples in a dataset, which consists of *Sensitivity* and *Specificity*. *Sensitivity* measures the degree to which positive samples are accurately classified, while *specificity* measures the degree to which negative samples are correctly classified. Its formula is as follows:

$$Gmean = \sqrt{Sensitivity \cdot Specificity} \tag{11}$$

Gmean is only high if the proportion of positive and negative classes correctly recognized is high. In this experiment, *Gmean* and accuracy are chosen to test the performance of unbalanced classification.

Software

ALL the calculations were implemented in Matlab 2018a and Python 3.11, and on an individual computer with an Intel(R) Core(TM) i7-8550U CPU @ 1.80 GHz 1.99 GHz and 16 GB of RAM memory, with the Windows 10 Professional operating system.

RESULTS AND DISCUSSION

Spectral comparative analysis

Figure 5a shows the NIR spectroscopy of DSP-contaminated mussel samples and healthy mussel samples mixed in the range of 950−1700 nm. Since the samples belong to the same species, the spectral curves tend to be similar. Figure 5b plots the average spectral curves of contaminated and healthy mussels, respectively, to show the spectral differences between the two groups of samples. In the wavelength range of 950 nm to 1100 nm, the spectral reflectance values of healthy samples are higher than those of DSP-contaminated samples. The reflection intensity of DSP-contaminated mussels is higher than that of healthy samples in 1150−1630 nm. In other wavelength ranges, the average spectral curves of the two samples almost overlap. This is because mussels will accumulate DSP toxins in soft tissues when they feed on toxin-producing algae. DSP toxins in contaminated mussels are in small quantities, and the DSP toxins barely have characteristic peaks in the infrared spectrum. Therefore, it is difficult to directly detect the amount of DSP toxin by the changes of spectral curve. Extremely complex chemical and enzymatic conversion mechanisms can occur in toxin-contaminated mussels (Liu et al., 2022c), resulting in

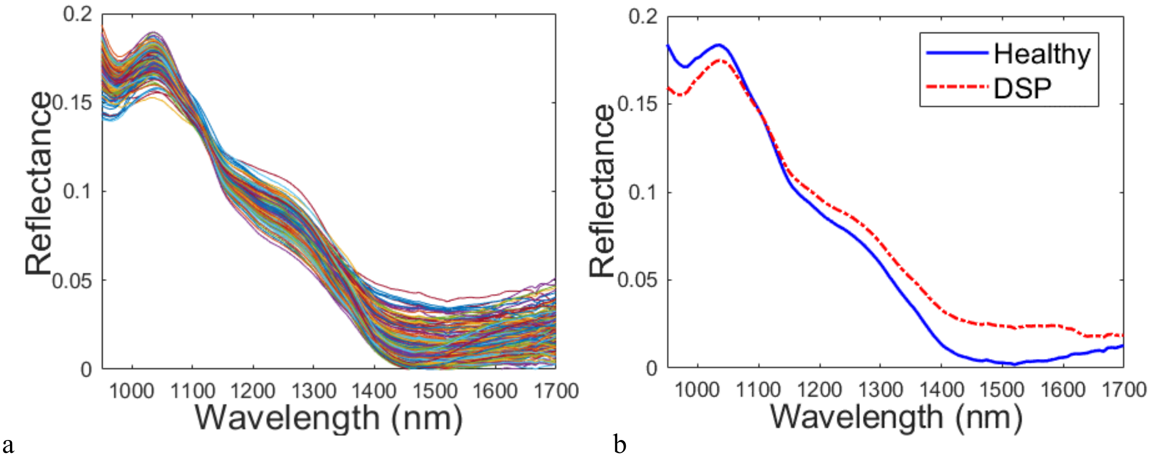

Fig. 5. Near-infrared spectrum of samples: a – original spectrogram, b – average spectrogram

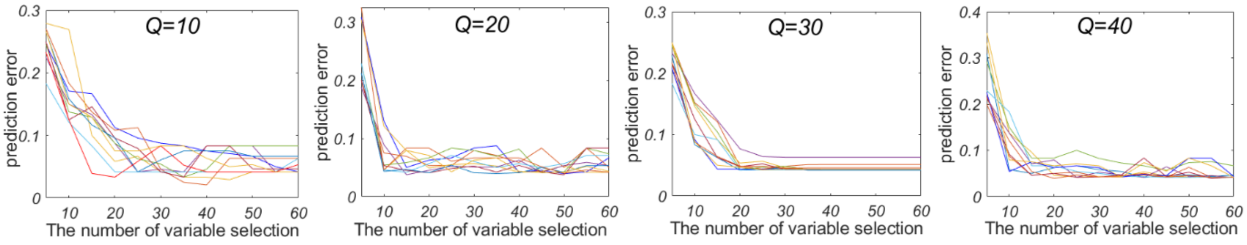

Fig. 6. Prediction error of the PHADIA algorithm changes with the number of variables

differences in the chemical composition of contaminated and healthy mussels.

In the near-infrared region, the absorption bands of many compounds have spectral signatures. In the near-infrared spectroscopy, the overtones and combinations of the fundamental vibrations of the O-H, N-H, C-H, and S-H functional groups are the most significant absorption bands. The chemical composition of the mussels changes after being polluted by the DSP toxin. These changes can produce molecular vibration information in the infrared spectral range. Spectral information on DSP toxin contamination can be obtained indirectly. From Figure 5, it can be seen that there are spectral differences between mussels contaminated with DSP toxins and healthy mussels in specific wavelengths, which is caused by compositional differences. These differences suggest that DSP-contaminated mussels can be detected using NIR spectroscopy.

Before building a classification model, the raw spectral data needs to be preprocessed to enhance the spectral features by making some corrections to it. Preprocessing algorithms include methods such as first derivative, second derivative, standard normal variable transformation, and quadrature signal correction (Jiang et al., 2016). In this study, orthogonal signal correction preprocessing is used to improve the accuracy of the classification model.

Determination of PHADIA algorithm parameters

The PHADIA method is used to screen the wavelength variable. For the mixed set spectral data of DSP mussels and healthy mussels after pretreatment, the parameters of PHADIA mainly included N and Q, and N is generally set to 10 000. The choice of Q value depends on several factors and optimized by cross-validation, and four values of Q, 10, 20, 30 and 40 are tested respectively. For each Q, the PHADIA algorithm is run 10 times and cross-validation is used to calculate the predictive model performance of SVM. Figure 6 shows the change in 5-fold cross-validation prediction error with the number of wavelength variables selected after running the PHADIA algorithm 10 times with different Q values. As shown in Figure 6, when Q takes different values, the prediction error is different after each operation of the PHADIA algorithm; at Q = 10, 20 and 40, the prediction wave fluctuates significantly. Considering the mean and variance of the prediction error, Q = 30 is optimal, and Q = 30 is selected for subsequent experimental variable screening.

Figure 7 shows the phase plot output by the PHADIA algorithm at Q = 30. It is divided into 4 areas, and

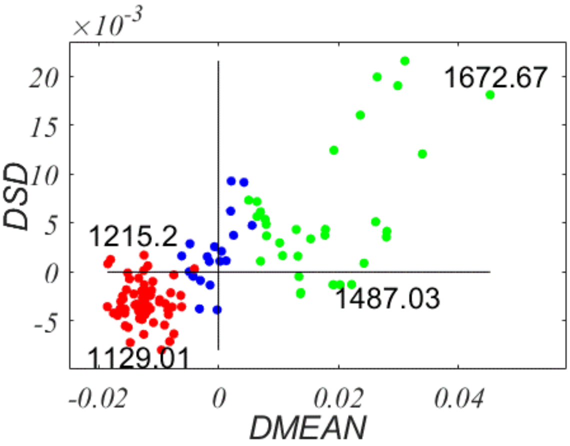

Fig. 7. Phase diagram of mixed datasets of DSP-contaminated and healthy mussel samples: green – information variables ($DMEAN > 0, p < 0.05$), blue – non-information variables ($p > 0.05$), red – the interfering variable ($DMEAN < 0$, $p < 0.05$)

intuitively divides the variables into information variables (green dots), non-information variables (blue dots) and interference variables (red dots). Phase 1 has the most informative wavelength variable. As shown in Figure 7, the wavelength variable 1672.67 nm stands out, indicating that this band has a high predictive value for DSP pollution and healthy mussel classification. Phase 2 contains fewer informative bands and is still being considered for modeling. These wavelength variables in phases 3 and 4 are likely to degrade the performance of the model, and they are eliminated as interfering variables. There are a total of 57 wavelength variables with *DMEAN* > 0 (phase 1 or phase 2 in Figure 7), of which 27 variables are significant ($p < 0.05$).

To illustrate the differences between the different types of variables, a wavelength variable from each of the four regions (phase 1: wavelength 1672.67 nm, phase 2: wavelength 1487.03 nm, phase 3: wavelength 1215.2 nm, phase 4: 1129.01 nm) is selected, and their prediction error distribution is shown in Figure 8. Taking the 1672.67 nm wavelength variable as an example, if it is included in the model, it can significantly reduce the prediction error, improve the stability of the prediction model and reduce the variance. The 1672.67 nm wavelength variable is thus a key band reflecting the physiological state of DSP toxins. In contrast, the variables in Figures 8a and 8c have poor predictive performance, because adding them to the model reduces the performance of the model.

To establish a streamlined DSP-contaminated mussel and healthy mussel mixed sample dataset classifier, all wavelength variables are sorted according to their *DMEAN* values, and a subset of 13 bands are selected using a forward strategy (from large to smallest). This is because the top 13 wavelength variables reach a minimum prediction error of 0.0462 in cross-validation, as shown in Figure 6. It can be seen that when the model input variables are less than 10, the 5-fold cross-validation error decreases significantly with the increase of variables, and the error is minimized when the variables are 13. As a result, thirteen wavelength

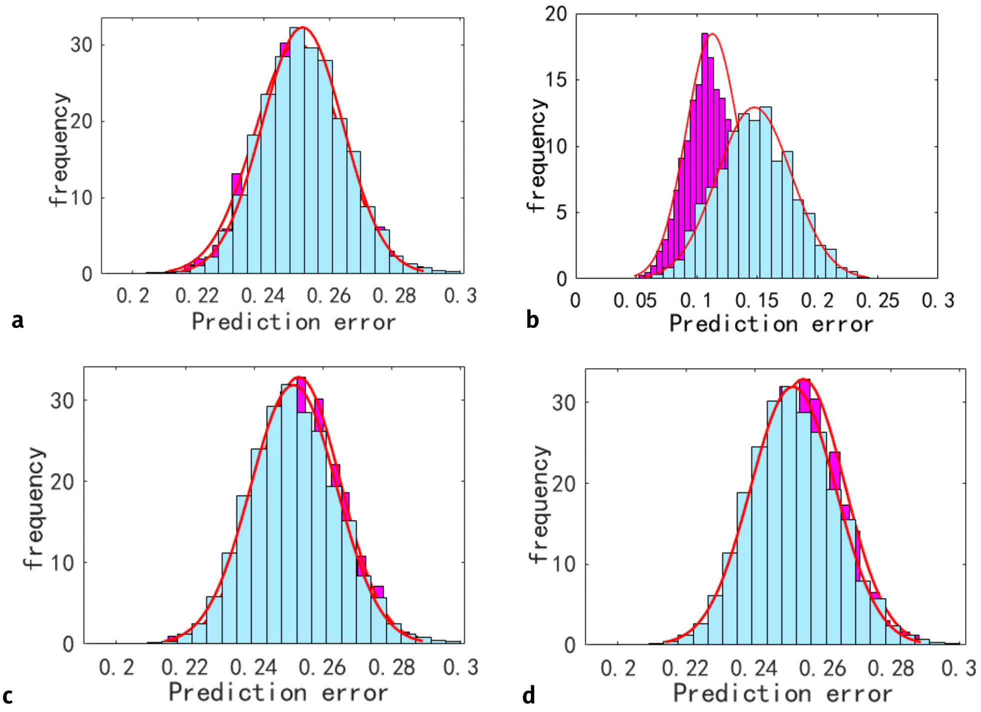

Fig. 8. Prediction error distribution of wavelength variables corresponding to four regions: a – wavelength 1215.2 nm, b – wavelength 1672.67 nm, c – wavelength 1129.01 nm, d – wavelength 1487.03 nm

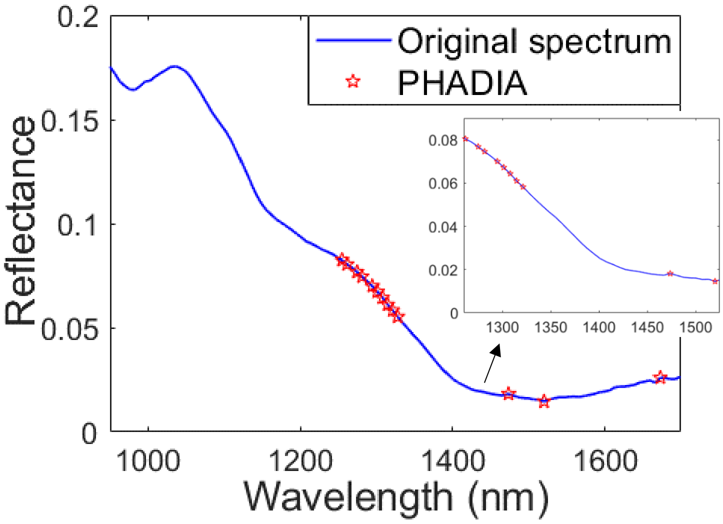

Fig. 9. Distribution of variables selected by the PHADIA algorithm

variables are finally identified for DSP-contaminated mussels and healthy mussel identification. Figure 9 shows the distribution of the characteristic bands for detecting DSP contamination on the spectral curve. The characteristic wavelengths screened by PHADIA algorithm are 1248.35 nm, 1254.98 nm, 1261.61 nm, 1274.87 nm, 1281.5 nm, 1294.76 nm, 1301.39 nm, 1308.02 nm, 1314.65 nm, 1321.28 nm, 1487.03 nm, 1520.18 nm, and 1672.67 nm respectively. According to the literature, protein-related bands were found at 1470 nm (Liu et al., 2022a), indicating that the band selected by the PHADIA algorithm can reflect spectral differences.

Comparison and analysis of detection performance of DSP toxin-contaminated mussels

This study used the PHADIA method to select feature variables and establish a DSP toxin detection model. DSP toxin detection is a problem of imbalance data classification in practice. To verify the superiority of the algorithm proposed in this paper for classifying unbalanced data, the IFSVM algorithm is compared with the classical algorithms, i.e. SVM and FSVM algorithms. All algorithms use 10-fold cross-verification. To reduce random effects, the algorithm is run ten times, and the final mean is taken as the final result. The experimental results are shown in Figure 10, which shows the changes in the *Gmean* and accuracy of the three classification algorithms with the proportion of healthy samples and DSP-contaminated samples in the training set, respectively. The test set includes 20 healthy samples and 20 DSP-contaminated samples. The training set remains unchanged for 100 healthy samples and reduces the number of DSP-contaminated samples from 100 to 10, 30 fewer at each time.

It can be seen from Figure 10 that the IFSVM method achieves the optimal effect in *Gmean* and accuracy in each unbalanced proportion of sample sets. Its mean value is greater than that of the FSVM and SVM methods, and the standard deviation is also smaller than that of the FSVM and SVM methods. When the imbalance ratio changes from 100:100 to 100:40, *Gmean* and accuracy do not decrease, but change non-regressively. The maximum *Gmean* and accuracy appear when the imbalance ratio is 100:40, which are 0.9886 and 98.83%, respectively. The results show that the IFSVM method performs better in dealing with unbalanced datasets. The *Gmean* and accuracy values of the SVM algorithm vary greatly with the proportion of healthy mussels and DSP-contaminated mussels, that is, the more unbalanced the training samples, the worse the detection effect. This is because the SVM algorithm does not consider the imbalance between healthy mussels and DSP-contaminated mussels, and obtains biased results. Although the FSVM algorithm takes into account the data imbalance, it only considers

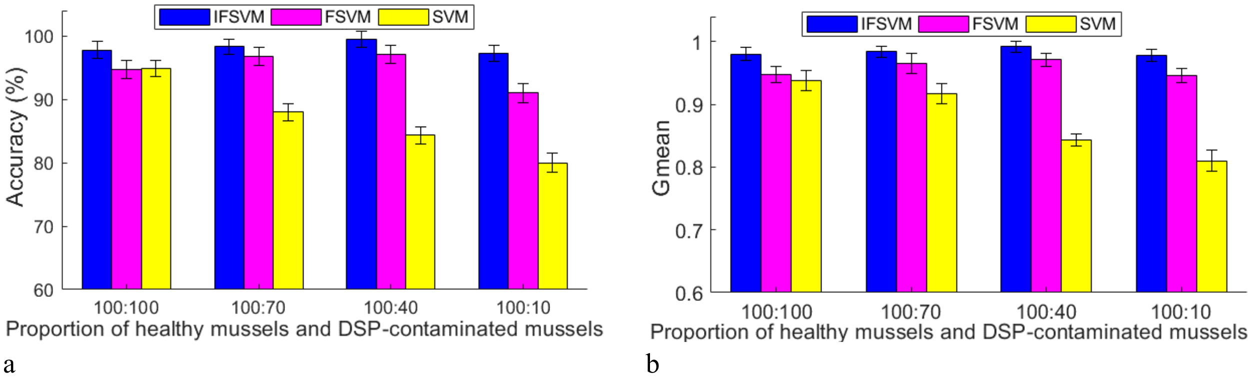

Fig. 10. *Gmean* and accuracy of different classification algorithms on data sets with different unbalanced proportions: a – the mean and standard deviation of the accuracy, b – the mean and standard deviation of Gmean

the distance from the sample to the center of the class. To this end, the algorithm cannot reflect the importance of the sample well on the unbalanced distribution of data. The IFSVM algorithm proposed in this paper not only considers the distance from the sample to the center of the class, but also considers the tightness around the sample and the amount of information of the sample when designing the fuzzy membership. It is basically not affected by the imbalance ratio, and the detection model proposed in this paper is suitable for unbalanced datasets.

Table 1. Optimal parameters of different classification algorithms on datasets with different unbalanced proportions

Dataset	Model	Parameters			
		C	*γ*	*α*	*M*
100: 100	SVM	2^{2}	2^{0}	–	–
	FSVM	2^{4}	2^{-1}	–	–
	IFSVM	2^{3}	2^{-3}	0.2	1.0
100: 70	SVM	2^{2}	2^{0}	–	–
	FSVM	2^{2}	2^{-1}	–	–
	IFSVM	2^{5}	2^{-11}	0.5	0.6
100: 40	SVM	2^{2}	2^{0}	–	–
	FSVM	2^{4}	2^{-1}	–	–
	IFSVM	2^{3}	2^{-3}	1.0	0.1
100: 10	SVM	2^{2}	2^{-9}	–	–
	FSVM	2^{4}	2^{-3}	–	–
	IFSVM	2^{1}	2^{-5}	0.3	0.7

However, when the classification accuracy is improved, the model parameters that need to be optimized by the proposed algorithm also increase. Compared with the SVM algorithm, although the IFSVM algorithm requires some additional calculation time when designing fuzzy membership, the computational complexity is theoretically comparable to that of SVM algorithm. Table 1 shows the optimal parameters of these algorithms on different unbalanced proportional datasets.

CONCLUSION

This study confirms that NIR spectroscopy is a fast, reliable and efficient method for the analysis of DSP pollution. The characteristic bands selected by the PHADIA algorithm can be effective in distinguishing healthy mussels and DSP-contaminated mussels. The IFSVM method can be used to solve the problem of unbalanced dataset classification in DSP toxin detection. The algorithm can effectively reduce the influence of unbalanced data on SVM and improve the accuracy of the classifier. Taking *Gmean* and classification accuracy as performance indicators, the performance of IFSVM is better than FSVM and SVM algorithms. The influence of PHADIA algorithm parameters and the parameters introduced by IFSVM model on detection performance is analyzed. The optimal parameters of the detection model are determined experimentally. However, it should be noted that the IFSVM algorithm not only improves the classification accuracy, but also needs to optimize the parameters. Therefore an effective parameter selection strategy should be further designed to improve the IFSVM algorithm.

ACKNOWLEDGMENT

The work was supported by the National Natural Science Foundation of China (62005109); Basic and Applied Basic Research Fund of Guangdong Province (2021A1515012440); Special fund for scientific and technological innovation strategy of Guangdong Province (pdjh2022a0312); Philosophy and Social Science Planning Project -Discipline Co-construction of Guangdong Province (No. GD22XJY32); Talent Program of Lingnan Normal University (ZL2049, ZL22023); 2022 Scientific Research Innovation Team Project of Lingnan Normal University (No. XJ2022000501).

REFERENCES

Bartlett, J. K., Maher, W. A., Purss, M. B. (2018). Near infra-red spectroscopy quantitative modelling of bivalve protein, lipid and glycogen composition using single-species versus multi-species calibration and validation sets. Spectrochim. Acta. A: Mol. Biomol. Spectr., 193, 537–557. https://doi.org/10.1016/j.saa.2017.12.046

Blanco, J., Mariño, C., Martín, H., Acosta, C. P. (2007). Anatomical distribution of diarrhetic shellfish poisoning (DSP) toxins in the mussel *Mytilus galloprovincialis*. Toxicon, 50(8), 1011–1018. https://doi.org/10.1016/j.toxicon.2007.09.002

Chen, J. Q., Wu H. Y., Yao L., Zheng G. C., Guo, M. M., Tan, Z. J., Zhai, Y. X., Mou, H. J. (2018). Rapid determination of diarrhetic shellfish poisoning in shellfish by colorimetric protein phosphatase inhibition assay based on immobilization of protein phosphatase 2A in sol-gel. Chinese J. Anal. Chem., 46(08), 1261–1268. http://doi.org/10.11895/j.issn.0253-3820.171529

Chen, X., Liu, K., Cai, J., Zhu, D., Chen, H. (2015). Identification of heavy metal-contaminated *Tegillarca granosa* using infrared spectroscopy. Anal. Meth., 7(5), 2172–2181. https://doi.org/10.1039/C4AY02396J

Corriere, M., Soliño, L., Costa, P. R. (2021). Effects of the marine biotoxins okadaic acid and dinophysistoxins on fish. J. Marine Sci. Eng., 9(3), 293. https://doi.org/10.3390/jmse9030293

Dirks, M., Poole, D. (2022). Automatic neural network hyperparameter optimization for extrapolation: Lessons learned from visible and near-infrared spectroscopy of mango fruit. Chem. Intel. Lab. Sys., 231, 104685. https://doi.org/10.1016/j.chemolab.2022.104685

Du, W., Zhang, X.H., Kong, L., Wang, D.N., Zheng, C.Y. (2020). Mouse biological method and ELISA method for the detection of diarrheal shellfish toxins. J. Zhejiang Agric. Sci., 61(10), 2132–2135. https://doi.org/10.16178/j.issn.0528-9017.20201052

Ganaie, M. A., Tanveer, M. (2021). Fuzzy least squares projection twin support vector machines for class imbalance learning. Appl. Soft Comp., 113(Part B), 107933. https://doi.org/10.1016/j.asoc.2021.107933

Guévélou, E., Allen Jr, S. K. (2016). Use of Near Infrared Reflectance Spectroscopy (NIRS) for the rapid compositional analysis of di-, tri-, and tetraploid eastern oysters (*Crassostrea virginica*). Aquaculture, 459, 203–209. https://doi.org/10.1016/j.aquaculture.2016.03.022

Guévélou, E., Carnegie, R. B., Whitefleet-Smith, L., Small, J. M., Allen Jr, S. K. (2021). near infrared reflectance spectroscopy to quantify *Perkinsus marinus* infecting *Crassostrea virginica*. Aquaculture, 533, 736063. https://doi.org/10.1016/j.aquaculture.2020.736063

Ioutsi, V. A., Panov, Y. M., Usol'tseva, L. O., Smolin, E. S., Antsupova, M. A., ..., N. G. Mokrysheva, N. G. (2022). Analysis of Serum Estrogens Using High-Performance Liquid Chromatography–Tandem Mass Spectrometry Coupled to Differential Ion Mobility Spectrometry. J. Anal. Chem. 77, 1760–1766. https://doi.org/10.1134/S1061934822140027

Jiang, W., Fang, J. L., Wang, S. W., Wang, R. T. (2016). Using CARS-SPA algorithm combined with hyperspectral to determine reducing sugars content in potatoes. J. North. Agric. Univ., 47(02), 88–95. https://doi.org/10.19720/j.cnki.issn.1005-9369.2016.02.013

Li, H. D., Xu, Q. S., Liang, Y. Z. (2017). A phase diagram for gene selection and disease classification. Chem. Intel. Lab. Syst., 167, 208–213. https://doi.org/10.1016/j.chemolab.2017.06.008

Lin, C. F., Wang, S. D. (2002). Fuzzy support vector machines. IEEE Trans. Neural Networks, 13(2), 464–471. https://doi.org/10.1109/72.991432

Liu, Y., Xu, L., Zeng, S., Qiao, F., Jiang, W., Xu, Z.(2022a). Rapid detection of mussels contaminated by heavy metals using near-infrared reflectance spectroscopy and a constrained difference extreme learning machine.

Spectrochim. Acta. Part A: Mol. Biomol. Spectrosc., 269, 120776. https://doi.org/10.1016/j.saa.2021.120776

Liu, Z. Y., Liu, Y., Qiao, F., Hao, B. L., Jiang, W., Xiong, J. F. (2022b). Rapid non-destructive detection of diarrheal shellfish poison in mussels based on near-infrared spectroscopy and multi-layer perceptron. Food Ferm. Ind., 1−10. https://doi.org/10.13995/j.cnki.11-1802/ts.032253

Liu, Y., Fu, Q., Xu, L., Wang, R., Jiang, W., Xu, Z. (2022c). Fast detection of diarrhetic shellfish poisoning toxins in mussels using NIR spectroscopy and improved twin support vector machines. Front. Marine Sci., 9, 907378. https://doi.org/10.3389/fmars.2022.907378

Mak, K. C., Yu, H., Choi, M. C., Shen, X., Lam, M. H., Martin, M., Lam, P. K. (2005). Okadaic acid, a causative toxin of diarrhetic shellfish poisoning, in green-lipped mussels *Perna viridis* from Hong Kong fishculture zones: Method development and monitoring. Marine Poll. Bull., 51(8–12), 1010–1017. https://doi.org/10.1016/j.marpolbul.2005.06.037

Manerio, E., Rodas, V. L., Costas, E., Hernandez, J. M. (2008). Shellfish consumption: A major risk factor for colorectal cancer. Med. Hypoth., 70(2), 409–412. https://doi.org/10.1016/j.mehy.2007.03.041

Melado-Herreros, Á., Nieto-Ortega, S., Olabarrieta, I., Ramilo-Fernández, G., Sotelo, C. G., Teixeira, B., Mendes, R. (2022). Comparison of three rapid non-destructive techniques coupled with a classifier to increase transparency in the seafood value chain: Bioelectrical impedance analysis (BIA), near-infrared spectroscopy (NIR) and time domain reflectometry (TDR). J. Food Eng., 322, 110979. https://doi.org/10.1016/j.jfoodeng.2022.110979

Raghavendra, A., Guru, D. S., Rao, M. K. (2021). Mango internal defect detection based on optimal wavelength selection method using NIR spectroscopy. Artificial Intel. Agric., 5, 43−51. https://doi.org/10.1016/j.aiia.2021.01.005

Savoia, S., Albera, A., Brugiapaglia, A., Di Stasio, L., Fergagina, A., Cecchinato, A., Bittante, G. (2020). Prediction of meat quality traits in the abattoir using portable and hand-held near-infrared spectrometers. Meat Sci., 16, 108017. https://doi.org/10.1016/j.meatsci.2019.108017

Srivastava, S., Mishra, H. N. (2022). Detection of insect damaged rice grains using visible and near infrared hyperspectral imaging technique. Chem. Intel. Lab. Syst., 221, 104489. https://doi.org/10.1016/j.chemolab.2021.104489

Wan, S., Li, Q., Yu, H., Liu, S., Kong, L. (2020). Estimating heritability for meat composition traits in the golden shell strain of Pacific oyster (*Crassostrea gigas*). Aquaculture, 516, 734532. https://doi.org/10.1016/j.aquaculture.2019.734532

Xiong, J. F., Qiao, F., Liu, Z. Y., Liu, Y., Hao, B. L., …, Lu, L. Q. (2022). Rapid and non-destructive detection for shellfish contaminated by heavy metal based on hyperspectral images. Environ. Eng., 40(10), 141−149. https://doi.org/10.13205/j.hjgc.202210019