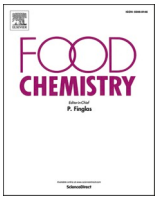
Food Chemistry 460 (2024) 140532

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem



Comparative evaluating laser ionization and iKnife coupled with rapid evaporative ionization mass spectrometry and machine learning for geographical authentication of *Larimichthys crocea*

Weibo Lu [a,1], Honghai Wang [a,1], Lijun Ge [a], Siwei Wang [c], Xixi Zeng [c], Zhujun Mao [c], Pingya Wang [f], Jingjing Liang [e], Jing Xue [a,**], Yiwei Cui [b,**], Qiaoling Zhao [f,**], Keyun Cheng [c,**], Qing Shen [c,d,*]

[a] *Collaborative Innovation Center of Seafood Deep Processing, Zhejiang Province Joint Key Laboratory of Aquatic Products Processing, Institute of Seafood, Zhejiang Gongshang University, Hangzhou, China*
[b] *College of Biology and Environmental Engineering, Zhejiang Shuren University, Hangzhou, China*
[c] *Panvascular Diseases Research Center, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou 324000, China*
[d] *Laboratory of Food Nutrition and Clinical Research, Institute of Seafood, Zhejiang Gongshang University, Hangzhou 310012, China.*
[e] *Zhejiang Provincial Institute for Food and Drug Control, Hangzhou 310052, China.*
[f] *Zhoushan Institute of Food & Drug Control, Zhoushan 316000, China.*

ARTICLE INFO

Keywords:
Larimichthys crocea
Traceability
Rapid evaporative ionization mass spectrometry
Machine learning

ABSTRACT

Larimichthys crocea (LYC) holds significant economic value as a marine fish species. However, inaccuracies in labeling its origin can adversely affect consumer interests. Herein, a laser assisted rapid evaporative ionization mass spectrometry (LA-REIMS) and machine learning (ML) was developed for geographical authentication. When compared to iKnife, the LA demonstrated to be superior owing to reduced thermal damage to sample tissue, enhanced automation, and ease of use. Analysis of LYC from six distinct geographical origins across China revealed a total of 798 ions, which were then subjected to six classifiers to establish ML models. Following hyperparameter optimization and feature engineering, the Chi2(15%)-KNN model exhibited the highest training and testing accuracy, achieving 98.4 ± 0.9% and 98.5 ± 1.4%, respectively. This LA-REIMS/ML methodology offers a rapid, accurate, and intelligent solution for tracing the origin of LYC, thereby providing valuable technical support for the establishment of traceability systems in the aquatic product industry.

1. Introduction

In recent years, the authenticity of geographical food origins has garnered increasing attention due to market globalization and recurrent food safety issues (Leal et al., 2015). With advancements in transportation and storage technologies, aquatic products, rich in omega-3 fatty acids, essential amino acids, and high-quality protein (Tacon & Metian, 2013), have become integral to the global food market (Kim et al., 2015). Among these, *Larimichthys crocea* (LYC), belonging to the Perciformes order and Sciaenidae family, is an economically significant marine fish species in East Asian countries, such as China, South Korea and Japan (Ao et al., 2015; Liu et al., 2020). It is highly prized by consumers for its delicious taste and superior nutritional value (Ma et al., 2021). However, variations in regional environments, farming methods, and genetic strains can significantly impact LYC's nutritional quality and flavor. These disparities result in price fluctuations across production regions, leading to concerns about potential economic fraud regarding food origin (Zheng et al., 2024). Hence, the development of authentication techniques for LYC's origin is paramount to safeguard its regional brand and product characteristics.

* Corresponding author at: Panvascular Diseases Research Center, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou 324000, China.
** Corresponding authors.
E-mail addresses: xuejing@zjgsu.edu.cn (J. Xue), ccdeyxzh@163.com (Y. Cui), zs_qlzhao@163.com (Q. Zhao), 1943882086@qq.com (K. Cheng), leonqshen@163.com (Q. Shen).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.foodchem.2024.140532
Received 3 April 2024; Received in revised form 29 June 2024; Accepted 18 July 2024
Available online 20 July 2024

Currently, a diverse array of analytical techniques is utilized in authenticating the geographical origin of food products, including DNA fingerprinting, nuclear magnetic resonance (NMR), near infrared spectroscopy (NIR), stable isotope analysis, and modern mass spectrometry technology (Amaral, 2021). For instance, Schütz et al. employed Fourier transform near infrared spectroscopy (FT-NIR) to successfully identify 101 milled grain maize samples originating from five distinct countries (Schütz et al., 2022). Dou et al. utilized gas chromatography–mass spectrometry (GC–MS) to determine the geographical origin of 161 camellia oil samples gathered from China's primary production regions, based on fatty acid profiles (Dou et al., 2024). Furthermore, Luo et al. employed multi-element and stable isotope analysis to discern the origins of 164 commodity crabs collected from eight different sites across China (Luo et al., 2020). Although these techniques offer remarkable accuracy in sample analysis, they often require intricate pre-treatment and extraction processes prior to sample introduction, coupled with sophisticated data processing procedures. This complexity poses significant challenges in achieving high-throughput analysis of aquatic products within the supply chain, thereby necessitating the exploration of alternative methods that can offer both accuracy and efficiency.

Rapid evaporative ionization mass spectrometry (REIMS) emerges as a pioneering ambient mass spectrometry technique, enabling swift in-situ analysis and authentication of food products (Black et al., 2017; Ross et al., 2021). For instance, Rigano et al. harnessed iKnife-REIMS technology to rapidly identify pistachio nuts from various geographical origins and varieties (Rigano et al., 2020). Liu et al. employed soldering iron-based REIMS to investigate the geographical traits of sorghum samples from China's major production areas (Liu et al., 2022). REIMS eliminates the need for sample preparation, streamlining data acquisition, analysis, and precise sample identification, all within a matter of seconds. This method involves direct ablation of biological tissues using sampling devices, followed by their introduction into the mass spectrometer for analysis (Barlow et al., 2021). Prior research has emphasized that aerosols from lipid-rich samples, such as aquatic products, harbor a wealth of lipid molecular information. When subjected to mass spectrometry, this information can yield distinct lipid fingerprint spectra, facilitating real-time identification (Shen et al., 2022). Conventionally, REIMS' front-end sampling devices include monopolar electric knives, bipolar forceps, electric soldering irons, and laser generators. However, devices like monopolar electric knives, bipolar forceps, and electric soldering irons generate aerosols through contact with conductive probes, limiting analysis to samples with suitable conductive and physical properties. This constraint, coupled with the need to replace or clean probes between samples, reduces analysis efficiency (Cameron et al., 2021). Additionally, these manual sampling devices introduce a risk of human error. Utilizing a laser generator as the sampling device for REIMS effectively addresses the aforementioned challenges. By scanning the sample with a laser beam, it generates aerosols without the need for direct contact (Genangeli et al., 2019). Cameron et al. demonstrated this capability by analyzing six commercially available cooking oils and three olive oils originating from protected production areas under European Union legislation using laser ablation-REIMS (LA-REIMS). The study achieved a significant separation of olive oils from three geographically protected Italian production areas, achieving 100% classification accuracy using random forest modeling (Cameron et al., 2021). This underscores the potential of LA-REIMS for high-throughput, automated, and accurate analysis.

After REIMS analysis, significant data volumes are typically generated, which are then processed through chemometric methods to extract crucial information, ultimately enhancing food authentication accuracy. Among these methods, machine learning has emerged as a promising approach in recent years. Its ability to integrate vast datasets, learn intricate relationships, and handle complex samples is noteworthy (Gredell et al., 2020). In comparison to traditional chemometric techniques like principal component analysis (PCA) and orthogonal partial least squares-discriminant analysis (OPLS-DA), machine learning demonstrates greater sensitivity to subtle data variations (Goyal et al., 2022). The integration of REIMS analysis with machine learning algorithms, such as support vector machine (SVM), k-nearest neighbor algorithm (KNN), naive Bayes (NB), and neural networks (NN), has shown remarkable success in authenticating food products with high accuracy (Song et al., 2024). However, the application of REIMS lipid fingerprinting coupled with machine learning techniques to determine the origins of aquatic products, especially LYC, remains uncommon. Further research in this area holds the potential to advance authenticity analysis for aquatic products and facilitate the implementation of quality, safety supervision, and traceability systems for these commodities.

In this study, a commercial diode blue laser generator was employed as a sampling device to pioneer a novel LA-REIMS methodology. To evaluate its performance, a comparative analysis was conducted against the traditional iKnife-REIMS approach. Notably, the integration of this technology with machine learning algorithms enabled the successful detection and precise classification of LYC from diverse geographical origins. This research offered a significant contribution, providing a valuable reference for automating the detection and expedited analysis of the authenticity of aquatic food products with regards to their geographic origin.

2. Materials and methods

2.1. Materials and reagents

Chromatographic-grade reagents, comprising methanol and acetonitrile, were procured from Merck (Darmstadt, Germany). Additionally, leucine enkephalin, serving as the internal standard with a purity of $\geq$97%, was purchased from Sigma-Aldrich (St. Louis, MO, USA). High-purity water, boasting a resistivity of 18.2 MΩ·cm at 25 °C, was sourced from the Millipore Milli-Q water system (Bedford, MA, USA). The LYC samples, with an average body weight of 450.0 ± 50.0 g, were procured from reliable local merchants in various regions across China, including Zhoushan, Taizhou, and Wenzhou in Zhejiang Province; Ningde in Fujian Province; and Weihai and Qingdao in Shandong Province. Over a two-year period, five batches of samples were collected annually from each of these six origins, with each batch consisting of six parallel samples. All samples, in their original physical state, were placed in polyethylene bags and stored in a freezer at −80 °C to ensure prompt testing. To mitigate the potential negative impacts of seasonal and hydrological variations on water quality parameters, including temperature, dissolved oxygen, pH, salinity, and light intensity, and to enhance the model's generalizability, the samples were procured annually for two consecutive years. Notably, a minimum interval of one month was maintained between each batch. For each batch, three LYC samples of comparable quality, originating from different production areas, were acquired as parallel specimens for rigorous testing. Before commencing the analysis, the frozen LYC samples were thawed at room temperature, followed by the careful removal of scales, skin, and internal organs. The remaining muscle tissue was delicately rinsed with high-purity water to ensure its purity. For testing, two rectangular pieces of dorsal muscle, each measuring 2 × 4 cm, were excised from both sides of the fish.

2.2. REIMS analysis

The chemical fingerprints of the samples were captured utilizing the REIMS analysis system from Waters Co., Ltd. in Beijing, China. This system featured a REIMS ionization source that was orthogonally mounted to the interface of a quadrupole time-of-flight (QTOF) mass spectrometer (Xevo G2-XS, Waters Co., Ltd., Milford, MA). For sample acquisition, a laser generator (K6 laser engraving machine, Shanghai DiaoTu Industrial Co., Ltd.) and an iKnife device (WSD151, Weller, Germany) were employed. The laser generator was affixed to a robotic arm, enabling precise adjustment of its position and angle to optimize

ionization efficiency. Operating at a wavelength of 450 nm and a power of 3 W, the laser possessed a spot diameter of 300 μm with a step size of 50 μm. The scanning area was set to a rectangular shape with dimensions of 8 × 4.5 mm, and the scanning rate was maintained at 25 $mm{\cdot}s^{-1}$, ensuring thorough and efficient sample analysis.

The iKnife device comprised a monopolar cutting mechanism equipped with a shortened knife blade of approximately 6 mm in length, operating in cutting mode at a power setting of 25 W. Each sampling cycle lasted 3 s, with the samples being cut repeatedly 5–7 times, interspersed with a 20-s interval between each cut. The resulting ionized aerosol was efficiently drawn into the REIMS source by a Venturi pump powered by 2 bar through a polytetrafluoroethylene (PTFE) tube. To mitigate background spectral interference, enhance signal intensity, and provide a reliable internal reference peak for lock-mass calibration, an auxiliary solvent was employed. Specifically, a 2-propanol solution containing 0.2 $ng{\cdot}\mu L^{-1}$ of leucine-enkephalin (m/z 554.2615) was injected into the REIMS source via a stainless-steel capillary (1/16″ outer diameter, 0.002″ inner diameter) at a flow rate of 0.1 $mL{\cdot}min^{-1}$. This solution mixed with the aerosol prior to entering the mass spectrometer. The REIMS analysis was conducted in negative ion mode, with the analysis range set to m/z 200 to 1000, ensuring comprehensive coverage of the desired chemical fingerprints.

2.3. Data analysis

After background subtraction and centering using MassLynx software (version 4.1, Waters, UK), the raw data of REIMS was exported in .txt format. The LYC samples collected from Weihai, Qingdao, Zhoushan, Taizhou, Wenzhou, and Ningde were designated as WH-IK, QD-IK, ZS-IK, TZ-IK, WZ-IK, and ND-IK for the iKnife-REIMS method, and WH-LA, QD-LA, ZS-LA, TZ-LA, WZ-LA, and ND-LA for the LA-REIMS method, based on the respective sampling technique. The suffixes 'IK' and 'LA' represent the iKnife-REIMS and LA-REIMS methodologies. In the processing of iKnife-REIMS data, the five scans (technical replicates) of the total ion current (TIC) were consolidated to generate a unique fingerprint for each sample. Conversely, for LA-REIMS data, 15-s detections in the TIC were accumulated to create the fingerprint for each sample. The structural identification of characteristic ions was achieved through a combined approach utilizing the LIPID MAPS prediction tool (http://www.lipidmaps.org/tools/index.html) and MS/MS analysis. The relative abundance of ions in the sample fingerprint was calculated via peak area normalization. Microsoft Excel software was employed to determine the mean value and standard deviation of the samples. TBtools software was used to create UpSet plots, while SIMCA-P 14.1 (Umetrics, Umea, Sweden) was utilized for unsupervised PCA and supervised OPLS-DA for multivariate statistical analysis. Additionally, SPSS 23.0 software was applied to perform linear discriminant analysis (LDA), validating the co-classification performance of multiple significantly different ions.

Based on LA-REIMS data, a classification model for LYC originating from diverse locations was developed using machine learning techniques in MATLAB 2022a (MathWorks Inc., Natick, USA). This process encompassed crucial steps such as model selection, optimization, feature engineering, and validation, as outlined by Cui et al. (2023). Six popular machine learning classifiers were chosen: decision trees (DT), discriminant analysis (DA), support vector machines (SVM), K-nearest neighbors (KNN), Naive Bayes (NB), and neural networks (NN). Before initiating model development, the dataset was randomly partitioned into a training set comprising 80% of the samples (6 sets, each with 24 fingerprint spectra) and a validation set containing the remaining 20% of the samples (6 sets, each with 6 fingerprint spectra). The model's establishment, refinement, and feature engineering were conducted through rigorous 10-fold cross-validation. The performance of the model was assessed using a confusion matrix and by calculating the model accuracy, as defined in Eq. (1). This evaluation process provided a comprehensive understanding of the model's discriminatory capabilities and its effectiveness in classifying LYC samples from different origins.

$$\text{Accuracy} = \frac{\text{All true classification}}{\text{All classification}} \times 100 \quad (1)$$

When discussing classification performance, "Accuracy" refers to the success rate of correctly identifying instances. "All true classification" signifies the instances that have been accurately classified across all classes, whereas "All classification" encompasses both accurately and inaccurately classified instances from all classes.

The detailed process of developing a machine learning model involved several key steps. Initially, a Bayesian optimizer was employed to model and scrutinize the optimization potential of six chosen classifiers: DT, DA, SVM, KNN, NB, and NN. After 30 iterations, the hyperparameters that resulted in the minimum classification error were deemed optimal for each model (Bischl et al., 2023). Subsequently, under these optimized hyperparameters, the performance of the models was gauged based on their accuracy on the training set, leading to the selection of the top three models with the highest accuracy for further examination. Moving forward, two from feature engineering - feature extraction and feature selection - were applied to diminish the feature dimensionality, thereby preventing model overfitting. Specifically, for feature extraction, the PCA function within the classification learner was leveraged for dimensionality reduction via PCA, focusing on extracting the most informative information based on variance. The number of principal components was chosen based on the criterion of capturing 95% or 99% of the total variance (Hasan & Abdulazeez, 2021). On the other hand, for feature selection, the Chi2 feature ranking algorithm was utilized by the classification learner. The accuracy and robustness of the resulting classification models were then assessed by retaining only the top 5%, 10%, 15%, and 20% of the features (Liu et al., 2019). Finally, taking into account the accuracy on both the training and testing sets as well as the training time, the optimal model for each classifier was determined. To address potential sampling bias, the original dataset underwent 10 random splits during the model optimization and validation process. A rigorous evaluation was then conducted through 50-fold cross-validation on each of the resulting 10 pairs of training and testing datasets. This comprehensive evaluation yielded the average accuracy and training time for each of the optimal models.

3. Results and discussion

3.1. Comparison of the sampling performance of iKnife-REIMS and LA-REIMS

3.1.1. Comparative analysis of thermal injury organization

Fig. 1 presented the macroscopic images of thermal injury in the muscle tissue of LYC after undergoing iKnife-REIMS and LA-REIMS. In Fig. 1A, the tissue injuries resulting from iKnife sampling exhibited variations in length, width, and depth, despite attempts to standardize the ionization time. Notably, the presence of prominent scorch marks at the sampling site edges indicated substantial impact on surrounding tissues. Conversely, Fig. 1B showcased the tissue injuries caused by laser sampling, which exhibited consistent sampling range and depth, with shallower wounds and minimal damage to the tissue surrounding the sampling point. The laser beam employed in this study inflicted less thermal damage to the tissue due to its narrower width compared to the iKnife blade. Additionally, the commercial laser generator had a lower ionization power, resulting in reduced tissue ablation. Consequently, the aerosol generated during LA-REIMS sampling was also diminished. These observations indicated that laser sampling may serve as a more precise and efficient method in practical applications (Genangeli et al., 2019).

3.1.2. Comparative analysis of signal stability and repeatability

The performance of the proposed method was rigorously evaluated in terms of accuracy and reproducibility, encompassing both intra-day

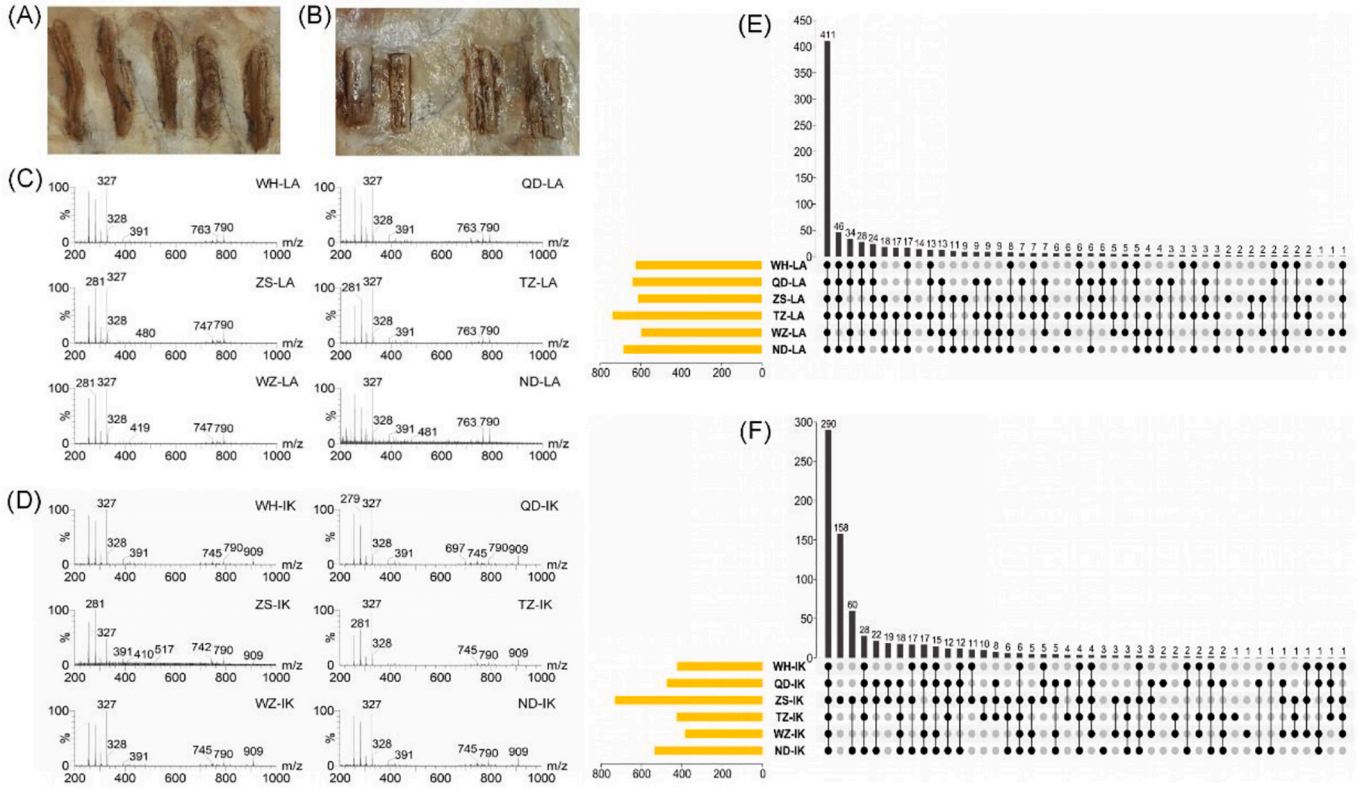


Fig. 1. Tissue damage after iknife-REIMS (A) and LA-REIMS (B) sampling ; representative fingerprints of LYC from different origins detected by LA -REIMS (C) and iKnife -REIMS (D); UpSet plot of lipid phenotypes of different LYC detected by (E) LA-REIMS and (F) iKnife-REIMS.

and inter-day precision. As a representative case, LYC samples from Taizhou were employed, with the characteristic ions *m/z* 327.2337, 391.2277, and 790.5444 serving as key indicators. For intra-day precision, errors were propagated from six duplicate ionizations of fish samples within the same batch, ensuring consistency. To assess inter-day precision, the samples were analyzed across three consecutive days. The findings were summarized in Table 1. The LA-REIMS intra-day accuracy (RSD) ranged from 3.08% to 7.71%, while the inter-day reproducibility hovered between 4.92% and 9.23%.The iKnife-REIMS intra-day precision (RSD) spanned from 3.70% to 9.01%, and its inter-day precision (RSD) ranged from 6.02% to 9.68%. Notably, LA-REIMS exhibited comparable intra-day and inter-day precision to iKnife-REIMS, with even lower values. This likely attributed to its automation, mechanization, and high controllability of sampling time, speed, and depth.

3.2. Analysis of lipid phenotypic differences

The characteristic REIMS fingerprints of LYC originating from Weihai, Qingdao, Zhoushan, Taizhou, Wenzhou, and Ningde, sampled using both laser and iKnife techniques, were depicted in Fig. 1C and D. Upon ionization by these two sampling devices, the lipid fingerprint profiles of LYC from various origins exhibited considerable overlap, indicating no significant differences in ion composition. However, minute variations in the abundances of some ions were observed. This could be attributed to the high degree of similarity in the composition of muscle lipids within the same fish species, whereas variations in growing waters, diets, and other factors might result in minor differences in muscle lipids.

The lipid ion phenotypes corresponding to all fingerprint profiles were consolidated and analyzed, yielding a substantial matrix dataset encompassing 800 feature ions (*m/z* bins) and 360 samples (2 sampling devices × 6 origins × 2 years × 5 batches × 3 parallels). The relative abundance of these feature ions was used for quantification. The Upset plot was employed to compare the lipid fingerprint profiles of all samples based on the detected *m/z* of ions. As depicted in Fig. 1E and F, LA-REIMS detected 798 features, accounting for 99.75% of the total, while iKnife-REIMS detected 773 features, representing 96.63% of the total. These results highlighted the high similarity in the lipid phenotype data of LYC from various origins obtained by both detection methods. Specifically, in LA-REIMS detection, there was an overlap of 411 ions in the lipid phenotype data of LYC from the six origins. Conversely, in iKnife-REIMS detection, 290 ions were common across LYC from all six locations. To gain further insights into the differences in lipid phenotypes at the molecular composition level among LYC from six distinct geographical origins, the features were sorted based on their relative abundance. Subsequently, the top 10 ions and their rankings in terms of relative abundance were summarized in Table S1 for each sample. In addition, the specific chemical structures of the identified features were determined. Notably, within the *m/z* range of 250–330 in Fig. 1C and D, fatty acid ion signals were prominently observed and identified as the $[FA-H]^-$ form (Cui et al., 2021). These $[FA-H]^-$ signals exhibited remarkably high abundance across all the lipid fingerprints of LYC. Among the LYC samples, *m/z* 327.2337, corresponding to $[FA22:6-H]^-$, showed the highest relative abundance, followed by *m/z* 281.2505

Table 1
Validation of intra-day and inter-day accuracy of iKnife-REIMS and LA-REIMS with Taizhou LYC samples.

Method	*m/z*	Intra-day accuracy		Inter-day	
		Abundance(%)	RSD	Abundance(%)	RSD
iknife-REIMS	327.2337	14.32 ± 1.29	9.01	15.09 ± 1.46	9.68
	391.2277	0.81 ± 0.03	3.70	0.83 ± 0.05	6.02
	790.5444	1.25 ± 0.08	6.40	1.12 ± 0.09	8.04
LA-REIMS	327.2337	12.45 ± 0.96	7.71	13.32 ± 1.23	9.23
	391.2277	1.30 ± 0.04	3.08	1.22 ± 0.06	4.92
	790.5444	1.59 ± 0.05	3.14	1.78 ± 0.10	5.62

([FA18:1 − H]$^-$), *m/z* 255.2339 ([FA16:0 − H]$^-$), and *m/z* 279.2338 ([FA18:2 − H]$^-$). This finding aligned with previous research by Ma et al., who analyzed the fatty acid composition of LYC using gas chromatography (Ma et al., 2021). However, a notable difference was that DHA (docosahexaenoic acid) exhibited the highest relative abundance in our study, while Ma et al. reported FA16:1, FA16:0, and FA18:2 as the top three fatty acids. This discrepancy could be attributed to the fact that the [FA − H]$^-$ signal in the REIMS lipid fingerprint may originate not only from the ionization of free fatty acids but also from the fragmentation of acyl chains in phospholipids during the ionization process. The DHA acyl chains in marine phospholipids are predominantly positioned at the sn-2 position of the glycerol backbone. This specific location renders them more prone to fragmentation during the REIMS ionization process, thereby generating [RCOO]$^-$ ions. This uneven fragmentation pattern is likely a significant contributor to the observed differential phenomenon (Cui et al., 2021; Song et al., 2020). Additionally, among the ions exhibiting higher abundance, PC, PE, and PI were detected. Their chemical structures were identified using the LIPIDMAPS search library and MS/MS analysis. For instance, considering the ion with *m/z* 790.5444, a preliminary inference based on the LIPIDMAPS search with a minimum mass difference of ±0.01 Da suggested it to be [PE40:6 − H]$^-$. The MS/MS spectrum of this ion (Fig. S1) revealed characteristic fragment ions such as *m/z* 279 ([$R_{18:2}$COO]$^-$), *m/z* 281 ([$R_{18:1}$COO]$^-$), *m/z* 283 ([$R_{18:0}$COO]$^-$), *m/z* 301 ([$R_{20:5}$COO]$^-$), *m/z* 303 ([$R_{20:4}$COO]$^-$), *m/z* 327 ([$R_{22:6}$COO]$^-$), and *m/z* 329 ([$R_{22:5}$COO]$^-$). These fragments enabled an initial determination of a partial fatty acyl structure. Consequently, it can be inferred that the ion at *m/z* 774 corresponds to combinations such as PE18:0/22:6, PE18:1/22:5, PE18:2/22:4, PE20:1/20:5, and PE20:2/20:4. An analysis of the top ten most abundant ions revealed that FA16:0, FA18:0, FA18:1, FA22:6, and their corresponding acyl chains exhibited higher abundances in LYC samples originating from all six tested locations, regardless of the specific lipid molecules. These findings were consistent with the typical lipid composition observed in marine fish species (Fernandes et al., 2014; Li et al., 2011).

In summary, the REIMS outcomes closely aligned with traditional mass spectrometry approaches in elucidating the lipid composition of LYC, demonstrating the proficiency of the LA-REIMS technology introduced in this study for achieving dependable lipid fingerprinting extraction results. Nevertheless, despite the ability of REIMS lipid fingerprints to distinguish LYC samples originating from various geographical locations, the fingerprints and associated datasets exhibited only subtle phenotypic variations. Consequently, to ensure precise qualitative classification of LYC samples, it was imperative to combine REIMS lipid fingerprinting with complementary analytical methodologies.

3.3. Multivariate statistical analysis

Multivariate statistical analysis was performed on the lipid fingerprint data of LYC samples obtained from the iKnife-REIMS and LA-REIMS.

3.3.1. PCA analysis

The chemometric comparison of lipid fingerprints acquired using two different sampling methods in LYC was conducted to assess their similarity. After computing the average relative abundance of lipid features in each sample group, the unique lipid fingerprint profiles for each set of samples were analyzed. Fig. S2 showcased the PCA score plot encompassing 12 sample sets derived from the two sampling approaches. Here, 23 principal components contributed to 74.7% of the total variance, with PC1 and PC2 accounting for 21.70% and 9.23%, respectively. Notably, samples from ZS-IK and ND-LA exhibited outlier tendencies, while the remaining sample groups displayed a higher degree of clustering. The clustering heatmap presented in Fig. S3 visually depicted the 12 lipid fingerprints. The normalized relative abundance of each lipid feature was represented by the color intensity of the heatmap rectangles, while the horizontal dendrogram highlighted the similarity and clustering patterns among the samples. Furthermore, to quantify the similarity between LA-REIMS and iKnife-REIMS technologies in generating lipid fingerprints, spectral similarity indices (SF) were employed. A SF value closer to 1 indicates a higher degree of similarity in mass spectra. As indicated in Table S2, all SF values exceeded 0.9000, suggesting a relatively minor difference between the detection data obtained using LA-REIMS and the traditional iKnife-REIMS method. Consequently, the novel LA-REIMS technology developed in this study exhibited comparable sampling performance to the established iKnife-REIMS approach.

Fig. 2A illustrated the PCA analysis of the iKnife-REIMS data. Here, the first seven principal components (PCs) cumulatively explained 53.2% of the variance, with PC1 to PC7 accounting for 26.30%, 11.20%, 4.20%, 3.39%, 3.05%, 2.72%, and 2.43% respectively. The score plot revealed that, apart from the ZS-IK samples exhibiting notable outlier tendencies, the remaining five samples exhibited a pronounced clustering pattern. This clustering poses challenges for PC1 and PC2 in effectively categorizing the samples. Fig. 2B presented the PCA analysis of LA-REIMS data obtained from LYC samples of diverse origins. The first eight PCs accounted for a cumulative variance of 52.2%, with PC1 to PC8 explaining 13.90%, 11.80%, 6.70%, 5.34%, 4.92%, 4.12%, 3.44%, and 1.89% respectively. In the LA-REIMS score plot (Fig. 2B), the high overlap of features shared by samples from various origins, totaling 411 (Fig. 1E), results in significant overlap among all six sample groups. Despite this overlap, samples originating from the same location still exhibited clustering tendencies, indicating the potential for intra-group classification.

The lipid fingerprint spectra of LYC samples originating from diverse geographical locations, when analyzed using both iKnife-REIMS and LA-REIMS techniques, displayed a remarkable degree of similarity. This similarity was quantitatively assessed through spectral similarity indices, which were summarized in Table S3 for iKnife-REIMS and Table S4 for LA-REIMS. Specifically, in the iKnife-REIMS analysis, the spectral similarity indices between ZS-IK and the other five LYC sample groups ranged from 0.8318 to 0.8803, while all other groups displayed indices exceeding 0.9. Similarly, LA-REIMS analysis consistently revealed fingerprint spectra similarity indices of LYC samples from various geographical origins to be >90%. These findings, coupled with PCA analysis, indicated a high degree of lipid phenotyping similarity among LYC samples regardless of their geographical origin.

3.3.2. OPLS-DA analysis

Supervised OPLS-DA was utilized to categorize the lipid phenotyping data of LYC samples acquired through two distinct sampling methods. In the OPLS-DA analysis of iKnife-REIMS-detected samples, the combination of 5 principal components (PCs) and 12 orthogonal components accounted for 84.9% of the cumulative variance, with PCs 1–5 individually explaining 19.7%, 18.8%, 17.4%, 16.0%, and 13.1% of the variance. Notably, the OPLS-DA model exhibited a significantly higher explained variance compared to the corresponding PCA model on the same dataset. This superiority is attributed to the orthogonal rotation of data projection employed by OPLS-DA, which enables accurate reclassification of initially scattered samples (i.e., those beyond the 95% confidence interval in PCA clustering) (Boccard & Rutledge, 2013). Consequently, the number of samples correctly assigned to their respective groups increased, enhancing the clustering accuracy of samples belonging to the same species. The score plot generated using the first two PCs (Fig. 2C) effectively showcased the distinct classification of QD-IK, ZS-IK, and TZ-IK sample groups, while a minor overlap was observed among the WH-IK, WZ-IK, and ND-IK groups. The permutation test with 100 iterations (Fig. 2D) yielded a $|R^2|$ value of 0.3864 and a $|Q^2|$ value of 0.4395, suggesting that the model was robust, with low overfitting risk and excellent predictive power.

In the OPLS-DA analysis of LA-REIMS-detected samples, five

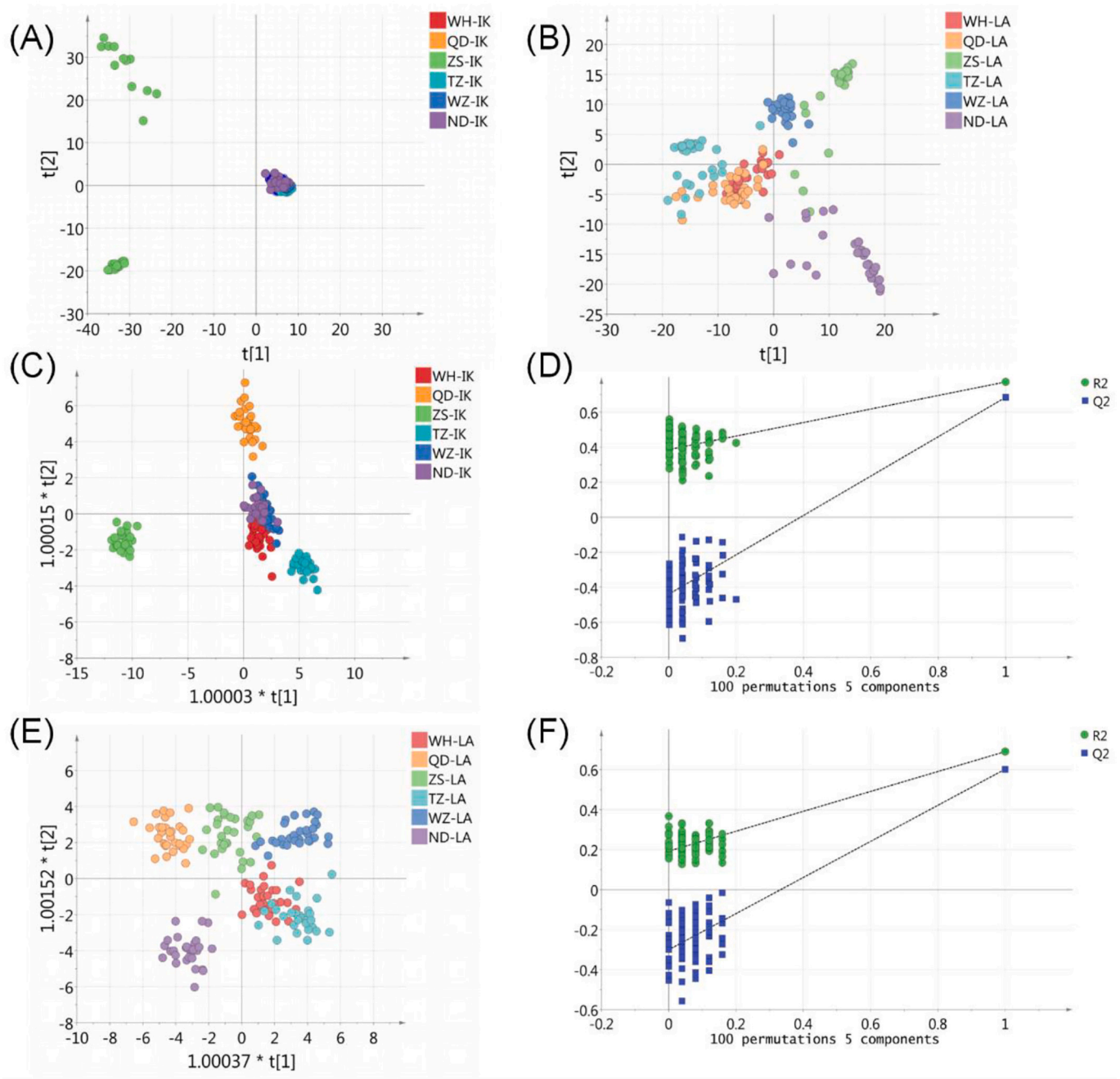


Fig. 2. PCA analysis of iKnife-REIMS (A) and LA-REIMS (B) on LYC from different origins; OPLS-DA analysis of LYC from different origins detected by iKnife-REIMS (C) and its substitution test (D), and OPLS-DA analysis of LYC from different origins detected by LA-REIMS (E) and its substitution test (F).

principal components (PCs) and seven orthogonal components collectively accounted for 82.7% of the cumulative variance. Specifically, PCs 1–5 explained 18.3%, 18.3%, 18.1%, 15.2%, and 12.7% of the variance, respectively. When examining the score plots (Fig. 2E) generated based on the first two PCs, it was observed that while the clustering of sample points within the same group was slightly more dispersed compared to Fig. 2B, the between-group differences were still discernible. Notably, only the WH-LA and TZ-LA groups exhibited a relatively higher degree of overlap. The permutation test conducted with 100 iterations (Fig. 2F) revealed a $|R^2|$ value of 0.1920 and a $|Q^2|$ value of 0.3005. These results indicated that the model was not overfitted and possesses a satisfactory predictive capability, similar to the findings from the previous analysis.

The aforementioned analysis revealed that the OPLS-DA models employed for both sampling methods achieved a model fit of <85% in classifying LYC samples based on their geographical origins. Notably, the score plots exhibited considerable overlap among sample points, indicating that while the OPLS-DA model surpassed PCA in performance, it still fell short of providing satisfactory results in distinguishing LYC samples from different geographical origins.

3.3.3. *Significant difference ion analysis*

Utilizing the OPLS-DA analysis, we identified ions with VIP values exceeding 1 as significant difference ions (Chen et al., 2023). Subsequently, we performed an analysis using the OPLS-DA model on the fingerprint spectra acquired through two distinct sampling methods, leading to the identification of 25 significant different ions for iKnife-REIMS and 29 for LA-REIMS. The corresponding *m/z* ratios and VIP values for these ions were compiled in Table 2.

Table 2
The VIP value of remarkable differential ions.

Measured ion (*m/z*)	Calculated Value	Mass Error (ppm)	TC: DB		VIP value		Characteristic MS2 ions	Comments
					iKnife-REIMS	LA-REIMS		
221.2391	221.2378	5.88	18:1	$[RCO-CO_2]^-$	–	2.2229		
222.1460	–	–	–	–	–	2.0984		Speculated to be the isotope peak of *m/z* 221.2391
223.2540	223.2535	2.24	18:0	$[RCO-CO_2]^-$	–	1.3455		
240.2076	–	–	–	–	–	1.6990		Unable to identify at present.
241.2175	241.2165	4.01	15:0	$[FA-H]^-$	–	1.0385		
253.2183	253.2165	6.95	16:1	$[FA-H]^-$	4.9042	2.6009		
255.2339	255.2322	6.63	16:0	$[FA-H]^-$	10.4480	11.6874		
256.2383	–	–	–	–	1.9553	2.4885		Speculated to be the isotope peak of *m/z* 255.2322
279.2338	279.2322	5.71	18:2	$[FA-H]^-$	11.3435	12.2440		
280.2345	–	–	–	–	2.0643	2.2607		Speculated to be the isotope peak of *m/z* 279.2322
281.2505	281.2478	9.47	18:1	$[FA-H]^-$	9.4495	11.0270		
282.2706	–	–	–	–	2.4181	2.1823		Speculated to be the isotope peak of *m/z* 281.2478
283.2657	283.2635	7.80	18:0	$[FA-H]^-$	3.7277	7.3512		
301.2192	301.2173	6.31	20:5	$[FA-H]^-$	3.6399	3.0464		
302.2171	–	–	–	–	1.6650	1.3094		Speculated to be the isotope peak of *m/z* 301.2173
303.2343	303.2330	4.25	20:4	$[FA-H]^-$	4.6325	6.0775		
304.2420	–	–	–	–	–	1.1730		Speculated to be the isotope peak of *m/z* 303.2330
309.2826	309.2799	8.71	20:1	$[FA-H]^-$	1.0547	–		
327.2337	327.2330	2.11	22:6	$[FA-H]^-$	15.9422	11.1693		
328.2510	–	–	–	–	4.8705	4.4345		Speculated to be the isotope peak of *m/z* 327.2337
329.2504	329.2486	5.47	22:5	$[FA-H]^-$	4.3699	4.1826		
330.2551	–	–	–	–	1.2168	–		Speculated to be the isotope peak of *m/z* 329.2504
337.3127	337.3112	4.33	22:1	$[FA-H]^-$	1.0585	–		
339.3291	339.3269	6.39	22:0	$[FA-H]^-$	–	1.0171		
367.3605	367.3582	6.29	24:0	$[FA-H]^-$	–	1.7797		
391.2277	391.2269	2.16	16:0	$[FA-H+C_3H_5O_4P]^-$	1.4975	1.4346		
417.2434	417.2425	2.06	18:1	$[FA-H+C_3H_5O_4P]^-$	1.3450	–		
419.2608	419.2582	6.14	18:0	$[FA-H+C_3H_5O_4P]^-$	1.0259	–		
715.4798	715.4791	0.98	38:1	$[PC-CHN(CH_3)_3]^-$	–	1.3901	253.2173, 255.2337, 281.2501, 283.2657, 309.2826, 715.4798	Composition of acyl chains: 18:0/18:1, 16:0/20:1, 16:1/20:0
742.5447	742.5392	7.41	36:2	$[PE-H]^-$	1.0447	–	279.2335, 281.2501, 283.2643, 417.2429, 742.5447	Composition of acyl chains: 18:0/18:2, 18:1/18:1
745.4481	745.445	4.10	38:6	$[PC-NH(CH_3)_3]^-$	1.4510	–	255.2337, 281.2501, 327.2337, 417.2429, 745.4481	Composition of acyl chains: 16:0/22:6, 18:1/20:5
747.4609	747.4606	0.34	38:5	$[PC-NH(CH_3)_3]^-$	–	1.1765	255.2337, 281.2501, 283.2643, 301.2183, 303.2337, 329.2486, 419.2593, 747.4609	Composition of acyl chains: 16:0/22:5, 18:0/20:5, 18:1/20:4
763.4806	763.4791	2.02	40:5	$[PC-CHN(CH_3)_3]^-$	–	2.1332	281.2501, 283.2643, 301.2183, 303.2337, 329.2486, 453.1976, 763.4806	Composition of acyl chains: 18:0/22:5, 18:1/22:4, 20:0/20:5, 20:1/20:4
764.5275	764.5236	5.12	38:5	$[PE-H]^-$	–	1.4606	253.2173, 255.2337, 279.2335, 281.2501, 283.2657, 301.2183, 303.2337, 307.2655, 329.2486, 764.5275	Composition of acyl chains: 16:0/22:5, 16:1/22:4, 18:0/20:5, 18:1/20:4, 18:2/20:3, 18:3/20:2
790.5444	790.5392	6.53	40:6	$[PE-H]^-$	1.0597	2.6948	279.2335, 281.2501, 283.2657, 301.2183, 303.2337, 327.2337, 329.2486, 790.5444	Composition of acyl chains: 18:0/22:6, 18:1/22:5, 18:2/22:4, 20:1/20:5, 20:2/20:4
792.5570	792.5549	2.65	40:5	$[PE-H]^-$	–	1.2723	281.2501, 283.2657, 301.2183, 303.2337, 309.2826, 329.2486, 792.5570	Composition of acyl chains: 18:0/22:5, 18:1/22:4, 20:0/20:5, 20:1/20:4
909.5501	909.5499	0.18	40:6	$[PI-H]^-$	3.2525	–	279.2335, 281.2501, 283.2657, 301.2183, 303.2337, 327.2337, 419.2591, 581.3165, 909.5501	Composition of acyl chains: 18:0/22:6, 18:1/22:5, 18:2/22:4, 20:1/20:5, 20:2/20:4
910.5552	–	–	–	–	1.4001	–		Speculated to be the isotope peak of *m/z* 909.5501

Note: TC: total carbon, DB: double bond. The comments in the column represent speculations on the structures of ions that have not been definitively identified, as well as the acyl chain compositions of identified phospholipid structures determined through MS/MS analysis.

3.3.4. LDA analysis

The collaborative classification potential of the significantly different ions obtained through both REIMS techniques was evaluated through LDA analysis. For the LDA model constructed using the significant ions from iKnife-REIMS, five canonical discriminant functions were formulated, collectively explaining 100% of the total variance. These functions contributed 74.5%, 13.3%, 6.0%, 4.4%, and 1.7% to the variance, respectively. When this model was applied to validate the classification of the original grouped samples, it achieved an accuracy rate of 95.0% (Table S5). Similarly, for the LDA model based on the significant ions detected by LA-REIMS, five canonical discriminant functions were established, collectively explaining 100% of the total

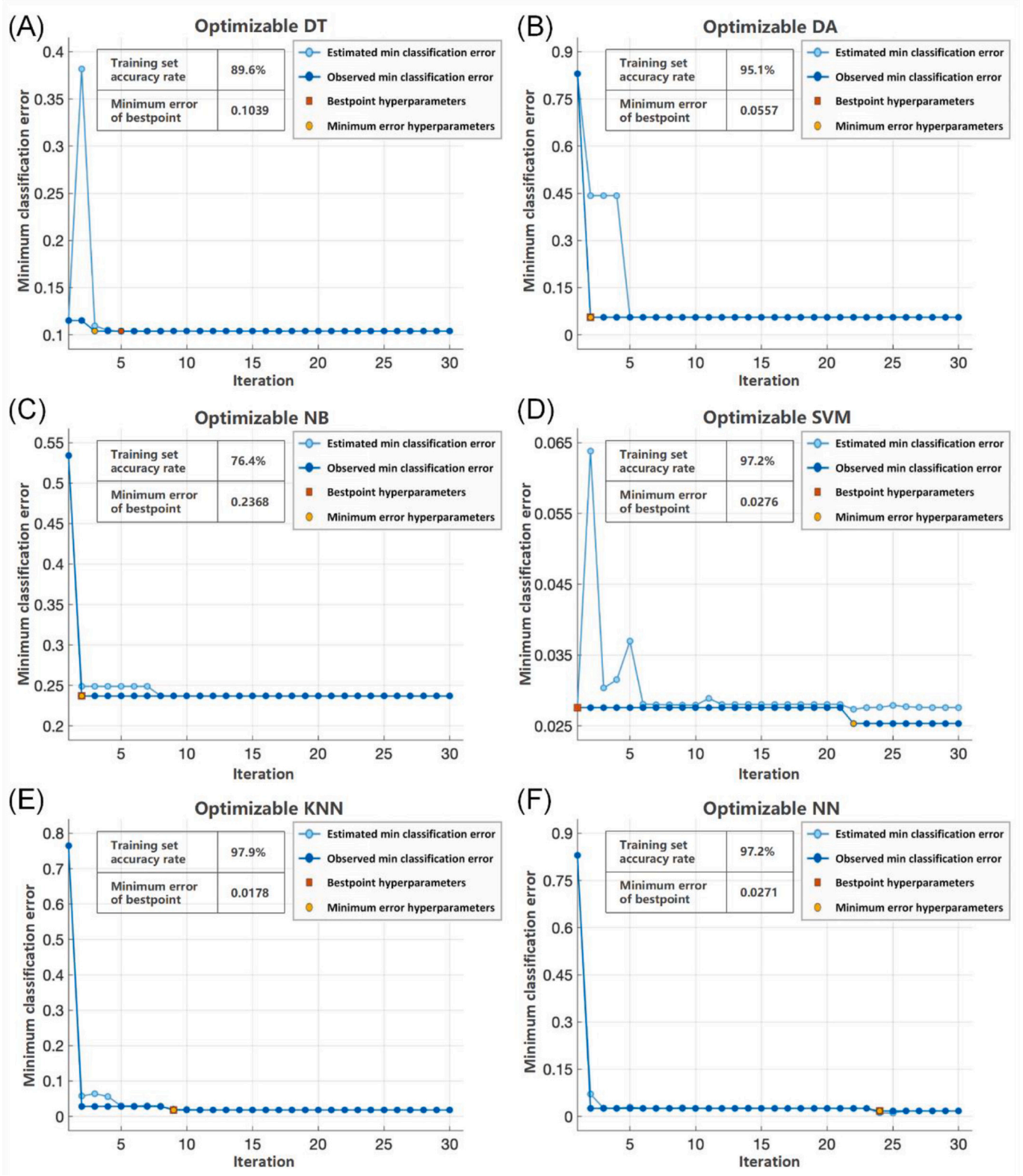


Fig. 3. The hyperparameters optimization of six machine learning classifiers: (A) DT, (B) DA, (C) NB, (D) SVM, (E) KNN and (F) NN classifiers.

variance. These functions contributed variances of 39.6%, 29.8%, 22.2%, 5.5%, and 3.0%, respectively. When this model was used to validate the original categorized samples, it achieved an accuracy rate of 97.8% (Table S6).

The results demonstrated that LDA models, when employed to classify LYC based on significant differential ions selected via OPLS-DA, can attain satisfactory classification accuracy. This consistency was observed regardless of whether the lipid phenotypic data originated from iKnife-REIMS or LA-REIMS sampling techniques. Nonetheless, this study entailed a significant amount of data analysis and relied on the LDA classification method, which necessitated the utilization of multiple multivariate statistical analysis models and extensive computational resources. In light of these challenges, there was a pressing need to develop chemometric techniques that offer greater convenience and a high level of automation, specifically for authenticity testing of food matrices that exhibit high similarity in lipid composition. Furthermore, a comparative analysis of the classification accuracies between the two LDA models revealed a superior accuracy for the lipid phenotypic data of LYC from various geographical origins when acquired through LA-REIMS. Consequently, lipid phenotypic data associated with this technique was chosen for subsequent analytical steps.

3.4. Machine learning classification model building and optimization

In prior research, experiments employing REIMS analysis paired with LiveID™ software have predominantly relied on principal component analysis-linear discriminant analysis (PCA-LDA) on PCA-reduced data to establish classification models (He et al., 2021; Mangraviti et al., 2021; Wang et al., 2019). PCA-LDA has proven its effectiveness in categorizing groups exhibiting significant variations in molecular fingerprint profiles. However, in scenarios where samples display a high degree of similarity in their molecular fingerprints or when multiple groups need to be classified concurrently within a single model, machine learning-based categorization models may outperform PCA-LDA (Gromski et al., 2015). Consequently, this study delved into comparing the classification capabilities of six machine learning classifiers for the LA-REIMS lipid fingerprint profile of LYC. Furthermore, it investigated the suitability of various feature engineering techniques in reducing the dimensionality of this dataset.

3.4.1. Optimal hyperparameters and optimal classifier selection

In this study, we selected six widely used machine learning classifiers in food authenticity applications to construct our models: DT, DA, SVM, KNN, NB, and NN. Prior to model development, we conducted thorough optimization procedures for each classifier. Fig. 3 outlined the optimization process and highlighted the optimal hyperparameters for each model. Subsequently, we successfully constructed six machine learning classification models using their respective optimal hyperparameters. The resulting training set accuracy rates for DT, DA, NB, SVM, KNN, and NN were 89.6% (Fig. 3A), 95.1% (Fig. 3B), 76.4% (Fig. 3C), 97.2% (Fig. 3D), 97.9% (Fig. 3E), and 97.2% (Fig. 3F), respectively.

3.4.2. Machine learning models optimization

The raw dataset used in this study comprised a 180 (sample size) × 800 (feature size) matrix, exhibiting a significant imbalance where the number of predictive factors (features) far exceeded the number of observations (samples). Such imbalanced datasets are commonplace in omics analyses and can pose challenges like extended training time and model overfitting when employing machine learning techniques for classification model development (Gromski et al., 2015). To ensure the credibility and scalability of our models, it was imperative to adopt a rational approach to reduce the feature count.

In this study, we compared two methods for dimensionality reduction: feature extraction and feature selection. The SVM, KNN, and NN models, constructed using optimized hyperparameters, were subjected to various techniques for reducing the data's dimensionality. The results, including the test and training accuracies as well as the training time of these three models, were summarized in Table 3.

Feature extraction was primarily achieved through PCA, a technique that diminished data dimensionality by transforming features into a reduced set of principal components. PCA accomplished this by capturing the covariation of feature variables, and the resulting principal components were then utilized as a new dataset, effectively reducing data dimensionality. While the new dataset's features differ from the original, they still encapsulate the fundamental data variations. In this context, PCA (95%) and PCA (99%) referred to the principal components that account for 95% and 99% of the variance, respectively. As evident in Table 3 and Fig. S4, following PCA-based dimensionality reduction, the SVM models employing PCA (95%)-SVM and PCA (99%)-SVM exhibited substantial reductions in model training time compared to the unreduced SVM model. Specifically, the PCA (95%)-SVM training time was 4.77 s, comprising only 8.4% of the pre-reduction training time. However, it was crucial to note that training and test set accuracies declined after PCA-based dimensionality reduction, with PCA (95%)-SVM exhibiting a more significant drop. This trend was also observed in the KNN (Table 3 and Fig. S5) and NN classifiers (Table 3 and Fig. S6), aligning with the findings of Gredell et al. (2019). The mass spectrometry data, devoid of missing values, exhibited high reliability following noise reduction and other data processing steps. However, the PCA dimensionality reduction process may inadvertently eliminate crucial information necessary for accurate sample classification, thereby reducing classification accuracy.

Another dimensionality reduction method employed in this study was feature selection, utilizing a supervised, recursive feature elimination approach to eliminate redundant features from the dataset. Among the various feature selection methods available, Filters, a pre-modeling data processing technique, emerged as a suitable choice when comparing different independent classifiers (Li et al., 2017). This method took into account feature dependencies, resulting in lower computational complexity and being agnostic to specific machine learning algorithms. Consequently, the features selected by this method could be effectively utilized across a range of classifiers. For data

Table 3
The accuracy of training and test sets and the training time of the SVM, KNN and NN model after data dimension reduction.

Model	Dimension reduction method	Training set accuracy rates (%)	Training time (s)	Test set accuracy rates (%)	Feature numbers
SVM	No dimension reduction	97.2	56.53	97.2	800/800
	PCA(95%)	91.7	4.77	88.9	15/143
	PCA(99%)	95.1	6.62	91.7	57/143
	Chi2(5%)	95.1	24.41	94.4	40/800
	Chi2(10%)	96.5	26.21	97.2	80/800
	Chi2(15%)	**97.2**	**26.06**	**100.0**	**120/800**
	Chi2(20%)	97.9	27.46	94.4	160/800
KNN	No dimension reduction	97.9	63.22	97.2	800/800
	PCA(95%)	91.7	3.40	86.1	15/143
	PCA(99%)	93.1	5.38	91.7	57/143
	Chi2(5%)	97.2	22.07	100.0	40/800
	Chi2(10%)	98.6	25.18	97.2	80/800
	Chi2(15%)	**99.3**	**28.88**	**100.0**	**120/800**
	Chi2(20%)	99.3	29.96	100.0	160/800
NN	No dimension reduction	97.2	59.94	97.2	800/800
	PCA(95%)	93.8	7.36	97.2	15/143
	PCA(99%)	96.5	10.62	97.2	57/143
	Chi2(5%)	**97.9**	**23.27**	**97.2**	**40/800**
	Chi2(10%)	97.9	28.92	94.4	80/800
	Chi2(15%)	98.6	38.28	97.2	120/800
	Chi2(20%)	97.9	58.24	94.4	160/800

dimensionality reduction, the chi-square test (Chi2) algorithm provided by MATLAB's classification learner was utilized. Chi2 is a commonly used supervised feature selection method in statistics (Tay & Shen, 2002). After ranking all features using the Chi2 algorithm, four variations of feature selection were implemented: selecting the top 5%, 10%, 15%, and 20% of features, denoted as Chi2(5%), Chi2(10%), Chi2 (15%), and Chi2(20%), respectively. Based on Figs. S4-S6, the high degree of similarity between samples posed a challenge during classification, potentially leading to the misclassification of LYC samples. No clear regularity was observed in this regard. However, analyzing Table 3 revealed that the training set accuracy of all models remained consistent, ranging from 95.1% to 99.3%, following Chi2 dimensionality reduction. Among these models, KNN's training set dimensionality reduction results exhibited the most stability. In terms of training time, it increased gradually as the number of selected features increased. Nonetheless, the Chi2 algorithm significantly reduced the training time of all other dimensionality reduction models by more than half, with exceptions being Chi2(15%)-NN and Chi2(20%)-NN.

The above conclusions collectively indicated that feature selection-based dimensionality reduction effectively improves the accuracy of the model and reduces training time.

3.4.3. Optimal machine learning model selection and validation

Based on the criteria of improved accuracy on both training and test sets compared to models without data dimension reduction, while minimizing training time, an optimal model was chosen for each classifier. After rigorous evaluation, three models emerged: Chi2(15%)-SVM, Chi2(15%)-KNN, and Chi2(5%)-NN. To mitigate sampling bias, 10 independent splits were performed on the training and testing subsets (Alakwaa et al., 2018). Additionally, 50-fold cross-validation was implemented to guard against overfitting and provide a more robust estimate of accuracy (Clark et al., 2020; Gredell et al., 2019). The summary of results in Table S7 revealed that all three models demonstrated excellent training set accuracies. Specifically, Chi2(15%)-KNN achieved the highest training set accuracy of 98.4 $\pm$ 0.9%, closely followed by Chi2(15%)-SVM (97.3 $\pm$ 0.8%) and Chi2(5%)-NN (96.2 $\pm$ 2.2%). Notably, these models also exhibited high accuracies on the testing set, with Chi2(15%)-KNN topping the list at 98.5 $\pm$ 1.4%, Chi2 (15%)-NN achieving 97.0 $\pm$ 1.4%, and Chi2(15%)-SVM scoring 96.4 $\pm$ 2.7%. A comparative analysis of the three optimal models revealed that Chi2(15%)-KNN significantly surpassed the other two in terms of accuracy, both on the training and testing sets. This finding underscores the effectiveness of Chi2(15%)-KNN in accurately discerning the geographical origins of LYC using LA-REIMS detection. Therefore, Chi2 (15%)-KNN stood as the most suitable model for this task. The study also involved testing an additional 60 samples (10 samples $\times$ 6 origins, purchased in different batches and not used for modeling) to evaluate the Chi2(15%)-KNN model's performance. The identification results were presented in Table S8. It could be observed that the classification rate achieved 96.7%.

The aforementioned facts clearly demonstrated that, in comparison to multivariate statistical analysis, machine learning can offer a superior approach for classifying and identifying LYC from various geographical origins. This method achieved higher accuracy in a faster, more convenient, and integrated manner. This underscored the vast potential of machine learning-guided REIMS pattern recognition technology in exploring and authenticating the geographical origins of aquatic food products.

4. Conclusions

This study has introduced a novel LA-REIMS combined with machine learning algorithms method that aimed at exploring the variances in lipid phenotypes of LYC sourced from different geographical regions and achieving an accurate identification of its geographic origin. Additionally, the performance of the developed LA-REIMS technology was compared with that of the traditional iKnife-REIMS technology for authenticating similar food matrices. Notably, LA-REIMS demonstrated comparable sampling performance to iKnife-REIMS. In contrast to the latter, LA-REIMS reduced tissue thermal damage, enhanced automation, and was well-suited for high-throughput analyses. To enhance classification accuracy, six classifiers were employed in the machine learning model based on LA-REIMS detection. After dimensionality reduction, the Chi2(15%)-KNN model exhibited the highest training and testing accuracies of 98.4 $\pm$ 0.9% and 98.5 $\pm$ 1.4%, respectively. This study's development of a novel LA-REIMS pattern recognition technology, guided by machine learning principles, enabled accurate, stable, and high-throughput analysis of samples with reduced thermal damage and increased automation. Despite the advantages of LA-REIMS, traditional analytical techniques, such as LC-MS, are still necessary in certain cases, such as when resolving structural isomers or when absolute quantitative analysis is required. We propose LA-REIMS as a screening tool for high-throughput analyses, eliminating the need for complex prior processing. This LA-REIMS/ML method presents a rapid and intelligent means of authenticating aquatic products, while providing technical support for establishing quality and safety supervision and traceability systems within the aquatic product industry.

CRediT authorship contribution statement

Weibo Lu: Writing – original draft, Investigation. **Honghai Wang:** Methodology, Formal analysis, Conceptualization. **Lijun Ge:** Data curation. **Siwei Wang:** Writing – review & editing. **Xixi Zeng:** Writing – review & editing. **Zhujun Mao:** Writing – review & editing. **Pingya Wang:** Methodology. **Jingjing Liang:** Validation, Resources. **Jing Xue:** Software. **Yiwei Cui:** Visualization, Validation, Conceptualization. **Qiaoling Zhao:** Supervision, Resources. **Keyun Cheng:** Writing – review & editing. **Qing Shen:** Writing – review & editing, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgement

The authors thank the National Natural Science Foundation of China (32172304), Zhejiang Provincial Public Welfare Technology Research Project (LTGN23C200007), State Administration for Market Regulation Science and Technology Project (2023MK059), Eyas Program Incubation Project of Zhejiang Provincial Administration for Market Regulation (CY2022232), and Zhejiang Provincial Drug Administration Science and Technology Plan (2021007).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2024.140532.

References

Alakwaa, F. M., Chaudhary, K., & Garmire, L. X. (2018). Deep learning accurately predicts estrogen receptor status in breast Cancer metabolomics data. *Journal of Proteome Research, 17*(1), 337–347.

Amaral, J. S. (2021). Target and non-target approaches for food authenticity and traceability. *Foods, 10(1), Article 1*.

Ao, J., Mu, Y., Xiang, L.-X., Fan, D., Feng, M., Zhang, S., Shi, Q., Zhu, L.-Y., Li, T., Ding, Y., Nie, L., Li, Q., Dong, W., Jiang, L., Sun, B., Zhang, X., Li, M., Zhang, H.-Q.,

Xie, S., & Chen, X. (2015). Genome sequencing of the perciform fish *Larimichthys crocea* provides insights into molecular and genetic mechanisms of stress adaptation. *PLoS Genetics, 11*(4), Article e1005118.
Barlow, R. S., Fitzgerald, A. G., Hughes, J. M., McMillan, K. E., Moore, S. C., Sikes, A. L., … Watkins, P. J. (2021). Rapid evaporative ionization mass spectrometry. *A Review on Its Application to the Red Meat Industry with an Australian Context. Metabolites, 11 (3), Article 3*.
Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery, 13*(2), Article e1484.
Black, C., Chevallier, O. P., Haughey, S. A., Balog, J., Stead, S., Pringle, S. D., … Elliott, C. T. (2017). A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry. *Metabolomics, 13*(12), 153.
Boccard, J., & Rutledge, D. N. (2013). A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock omics data fusion. *Analytica Chimica Acta, 769*, 30–39.
Cameron, S. J. S., Perdones-Montero, A., Van Meulebroek, L., Burke, A., Alexander-Hardiman, K., Simon, D., … Takáts, Z. (2021). Sample preparation free mass spectrometry using laser-assisted rapid evaporative ionization mass spectrometry: Applications to microbiology, metabolic biofluid phenotyping, and food authenticity. *Journal of the American Society for Mass Spectrometry, 32*(6), 1393–1401.
Chen, J.-N., Zhang, Y.-Y., Huang, X., Wang, H.-P., Dong, X., Zhu, B., & Qin, L. (2023). Analysis of lipid molecule profiling and conversion pathway in mandarin fish (Siniperca chuatsi) during fermentation via untargeted Lipidomics. *Journal of Agricultural and Food Chemistry, 71*(22), 8673–8684.
Clark, S. L., Hattab, M. W., Chan, R. F., Shabalin, A. A., Han, L. K. M., Zhao, M., … van den Oord, E. J. C. G. (2020). A methylation study of long-term depression risk. *Molecular Psychiatry, 25(6), Article 6*.
Cui, Y., Lu, W., Xue, J., Ge, L., Yin, X., Jian, S., Li, H., Zhu, B., Dai, Z., & Shen, Q. (2023). Machine learning-guided REIMS pattern recognition of non-dairy cream, milk fat cream and whipping cream for fraudulence identification. *Food Chemistry, 429*, Article 136986.
Cui, Y., Wang, H., Zhao, Q., Zhu, X., Wang, P., Xue, J., Chen, K., & Shen, Q. (2021). Real-time detection of authenticity and adulteration of krill phospholipids with soybean phospholipids using rapid evaporative ionization mass spectrometry: Application on commercial samples. *Food Control, 121*, Article 107680.
Dou, X., Wang, X., Ma, F., Yu, L., Mao, J., Jiang, J., Zhang, L., & Li, P. (2024). Geographical origin identification of camellia oil based on fatty acid profiles combined with one-class classification. *Food Chemistry, 433*, Article 137306.
Fernandes, C. E., Da Vasconcelos, M. A. S., De Almeida Ribeiro, M., Sarubbo, L. A., Andrade, S. A. C., & De Filho, A. B. M. (2014). Nutritional and lipid profiles in marine fish species from Brazil. *Food Chemistry, 160*, 67–71.
Genangeli, M., Heeren, R. M. A., & Porta Siegel, T. (2019). Tissue classification by rapid evaporative ionization mass spectrometry (REIMS): Comparison between a diathermic knife and CO_2 laser sampling on classification performance. *Analytical and Bioanalytical Chemistry, 411*(30), 7943–7955.
Goyal, K., Kumar, P., & Verma, K. (2022). Food adulteration detection using artificial intelligence: A systematic review. *Archives of Computational Methods in Engineering, 29*(1), 397–426.
Gredell, D. A., Schroeder, A. R., Belk, K. E., Broeckling, C. D., Heuberger, A. L., Kim, S.-Y., … Wheeler, T. L. (2020). Comparison of machine learning algorithms for predictive modeling of beef attributes using rapid evaporative ionization mass spectrometry (REIMS) data*. In *Mass spectrometry imaging in food analysis*. CRC Press.
Gredell, D. A., Schroeder, A. R., Belk, K. E., Broeckling, C. D., Heuberger, A. L., Kim, S.-Y., … Prenni, J. E. (2019). Comparison of machine learning algorithms for predictive modeling of beef attributes using rapid evaporative ionization mass spectrometry (REIMS) data. *Scientific Reports, 9(1), Article 1*.
Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis – A marriage of convenience or a shotgun wedding. *Analytica Chimica Acta, 879*, 10–23.
Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining, 2 (1), Article 1*.
He, Q., Yang, M., Chen, X., Yan, X., Li, Y., He, M., Liu, T., Chen, F., & Zhang, F. (2021). Differentiation between Fresh and Frozen–Thawed Meat using Rapid Evaporative Ionization Mass Spectrometry: The Case of Beef Muscle. *Journal of Agricultural and Food Chemistry, 69*(20), 5709–5724.
Kim, H., Suresh Kumar, K., & Shin, K.-H. (2015). Applicability of stable C and N isotope analysis in inferring the geographical origin and authentication of commercial fish (mackerel, yellow croaker and Pollock). *Food Chemistry, 172*, 523–527.
Leal, M. C., Pimentel, T., Ricardo, F., Rosa, R., & Calado, R. (2015). Seafood traceability: Current needs, available tools, and biotechnological challenges for origin certification. *Trends in Biotechnology, 33*(6), 331–336.
Li, G., Sinclair, A. J., & Li, D. (2011). Comparison of lipid content and fatty acid composition in the edible meat of wild and cultured freshwater and marine fish and shrimps from China. *Journal of Agricultural and Food Chemistry, 59*(5), 1871–1881.
Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys, 50*(6), 94:1–94:45.
Liu, H., Zhou, M., & Liu, Q. (2019). An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica, 6*(3), 703–715.
Liu, Q., Lin, H., Chen, J., Ma, J., Liu, R., & Ding, S. (2020). Genetic variation and population genetic structure of the large yellow croaker (*Larimichthys crocea*) based on genome-wide single nucleotide polymorphisms in farmed and wild populations. *Fisheries Research, 232*, Article 105718.
Liu, T., Wang, W., He, M., Chen, F., Liu, J., Yang, M., Guo, W., & Zhang, F. (2022). Real-time traceability of sorghum origin by soldering iron-based rapid evaporative ionization mass spectrometry and chemometrics. *ELECTROPHORESIS, 43*(18–19), 1841–1849.
Luo, R., Jiang, T., Chen, X., Zheng, C., Liu, H., & Yang, J. (2019). Determination of geographic origin of Chinese mitten crab (Eriocheir sinensis) using integrated stable isotope and multi-element analyses. Food Chemistry, 274, 1–7.Ma, R., Meng, Y., Zhang, W., & Mai, K. (2020). Comparative study on the organoleptic quality of wild and farmed large yellow croaker *Larimichthys crocea*. *Journal of Oceanology and Limnology, 38*(1), 260–274.
Ma, X., Mei, J., & Xie, J. (2021). Effects of multi-frequency ultrasound on the freezing rates, quality properties and structural characteristics of cultured large yellow croaker (*Larimichthys crocea*). *Ultrasonics Sonochemistry, 76*, Article 105657.
Mangraviti, D., Rigano, F., Arigò, A., Dugo, P., & Mondello, L. (2021). Differentiation of Italian extra virgin olive oils by rapid evaporative ionization mass spectrometry. *LWT, 138*, Article 110715.
Rigano, F., Stead, S., Mangraviti, D., Jandova, R., Petit, D., Marino, N., & Mondello, L. (2020). Use of an "Intelligent Knife". In *(iknife), Based on the Rapid Evaporative Ionization Mass Spectrometry Technology, for Authenticity Assessment of Pistachio Samples *. In Mass Spectrometry Imaging in Food Analysis*. CRC Press.
Ross, A., Brunius, C., Chevallier, O., Dervilly, G., Elliott, C., Guitton, Y., … Vanhaecke, L. (2021). Making complex measurements of meat composition fast: Application of rapid evaporative ionisation mass spectrometry to measuring meat quality and fraud. *Meat Science, 181*, Article 108333.
Schütz, D., Riedl, J., Achten, E., & Fischer, M. (2022). Fourier-transform near-infrared spectroscopy as a fast screening tool for the verification of the geographical origin of grain maize (*Zea mays L.*). *Food Control, 136*, Article 108892.
Shen, Q., Lu, W., Cui, Y., Ge, L., Li, Y., Wang, S., Wang, P., Zhao, Q., Wang, H., & Chen, J. (2022). Detection of fish frauds (basa catfish and sole fish) via iKnife rapid evaporative ionization mass spectrometry: An in situ and real-time analytical method. *Food Control, 142*, Article 109248.
Song, G., Guo, X., Li, Q., Lv, J., Wang, D., Yuan, T., Liu, S., Li, L., Liao, J., Zhang, M., Shen, Q., Zheng, F., & Gong, J. (2024). Mislabeling identification of fresh retail beef cuts using machine learning – Guided REIMS lipidomic fingerprints. *Food Control, 161*, Article 110401.
Song, G., Wang, H., Zhang, M., Zhang, Y., Wang, H., Yu, X., Wang, J., & Shen, Q. (2020). Real-time monitoring of the oxidation characteristics of Antarctic krill oil (*Euphausia superba*) during storage by electric soldering Iron ionization mass spectrometry-based Lipidomics. *Journal of Agricultural and Food Chemistry, 68*(5), 1457–1467.
Tacon, A. G. J., & Metian, M. (2013). Fish matters: Importance of aquatic foods in human nutrition and global food supply. *Reviews in Fisheries Science, 21*(1), 22–38.
Tay, F. E. H., & Shen, L. (2002). A modified Chi2 algorithm for discretization. *IEEE Transactions on Knowledge and Data Engineering, 14*(3), 666–670.
Wang, H., Cao, X., Han, T., Pei, H., Ren, H., & Stead, S. (2019). A novel methodology for real-time identification of the botanical origins and adulteration of honey by rapid evaporative ionization mass spectrometry. *Food Control, 106*, Article 106753.