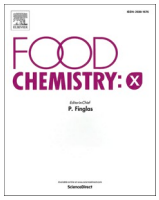
Food Chemistry: X 22 (2024) 101412

Contents lists available at ScienceDirect

Food Chemistry: X

journal homepage: www.sciencedirect.com/journal/food-chemistry-x



Geographical identification of Chinese wine based on chemometrics combined with mineral elements, volatile components and untargeted metabonomics

Kexiang Chen, Hongtu Xue, Qi Shi, Fan Zhang, Qianyun Ma, Jianfeng Sun, Yaqiong Liu, Yiwei Tang, Wenxiu Wang *

College of Food Science and Technology, Hebei Agricultural University, Baoding, Hebei 071000, China

ARTICLE INFO

Keywords:
Wine
Geographical origin
Mineral element
Volatile component
Untargeted metabolomics
Machine learning algorithms

ABSTRACT

Identifying the geographic origin of a wine is of great importance, as origin fakery is commonplace in the wine industry. This study analyzed the mineral elements, volatile components, and metabolites in wine using inductively coupled plasma-mass spectrometry, headspace solid phase microextraction gas chromatography–mass spectrometry, and ultra-high-performance liquid chromatography-quadrupole-exactive orbitrap mass spectrometry. The most critical variables (5 mineral elements, 13 volatile components, and 51 metabolites) for wine origin classification were selected via principal component analysis and orthogonal partial least squares discriminant analysis. Subsequently, three algorithms—K-nearest neighbors, support vector machine, and random forest —were used to model single and fused datasets for origin identification. These results indicated that fused datasets, based on feature variables (mineral elements, volatile components, and metabolites), achieved the best performance, with predictive rates of 100% for all three algorithms. This study demonstrates the effectiveness of a multi-source data fusion strategy for authenticity identification of Chinese wine.

1. Introduction

Wine, a fermented beverage made from fresh grapes or grape juice, undergoes total or partial fermentation and contains a specific alcohol level. Its complex composition includes water, ethanol, sugar, glycerol, organic acids, phenols, mineral elements, vitamins, and volatile compounds (Snopek et al., 2018). The quality of wine is affected by various factors, such as soil, climate, and water source, making its origin a crucial determinant of its characteristics (Marchionni et al., 2013). Many countries are renowned for producing high-quality wines. In China, societal development and increasing consumer demand, coupled with the growing popularity of wine culture, have led to a substantial expansion of the wine market. The trend of wine consumption and production in China is markedly positive. In 2022, China consumed approximately 880 million liters of wine (accounting for 4% of the global total) and produced 420 million liters (1.6% of the global total), ranking eighth and twelfth globally, respectively (OIV, 2022). High-quality wines, being of remarkable economic value, have unfortunately attracted some illegal traders who deceive consumers by falsifying geographical labels. This has resulted in the market being flooded with innumerable low-quality and overpriced wines. Therefore, verifying the geographical source of wines is increasingly crucial to protect high-value geographical indication products and prevent brand reputation damage and consumer interests.

Analytical techniques to evaluate the authenticity of wine-producing areas are increasingly utilized. Mineral elements have been established as remarkable chemical indicators of the local geographical environment (Gao et al., 2022; Plotka-Wasylka, Frankowski, Simeonov, Polkowska, & Namiesnik, 2018). When geochemical or soil data between regions are too similar, additional parameters such as the analysis of volatile compounds or metabolites can aid in further differentiation (Majchrzak, Wojnowski, & Plotka-Wasylka, 2018). Headspace solid-phase microextraction gas chromatography–mass spectrometry (HS-SPME-GC–MS) is mainly used for analyzing volatile components, allowing for the identification of chromatographic peaks and the acquisition of relative quantitative information without standard samples (Liang, Xie, & Chan, 2004). Non-targeted metabonomics, a method that analyzes many metabolites produced under specific conditions, can

* Corresponding author.
E-mail address: cauwwx@hebau.edu.cn (W. Wang).

https://doi.org/10.1016/j.fochx.2024.101412
Received 20 February 2024; Received in revised form 7 April 2024; Accepted 22 April 2024
Available online 25 April 2024

identify reliable markers indicative of different geographical origins (Cao, Du, Tang, Xi, & Chen, 2021). Insights from prior research indicate that single analytical technologies have some limitations in the accuracy of origin classification. To improve wine classification and certification, integrating data from various technologies to explore collaborative and complementary information can enhance the model's classification and prediction capabilities. Compared to relying on data from a single technology, merging data from complementary technologies provides accurate information about the samples, leading to improved inferences (classification with a low error rate and predictions with a low degree of uncertainty), feasible food sample differentiation, and enhanced authenticity discrimination. Mir-Cerdà et al. (2022) demonstrated that combining biogenic amine and mineral element data led to an increasingly comprehensive model and superior classification outcomes compared to the results obtained using a single type of data. Other studies have shown that fusing data in food certification can achieve high classification accuracy compared to single data types (Drivelos, Higgins, Kalivas, Haroutounian, & Georgiou, 2014; Longobardi, Casiello, Sacco, Tedone, & Sacco, 2011). However, reports on tracing wine origins through the combined analysis of mineral elements, volatile compounds, and metabolites are limited.

The purpose of this study is to comprehensively characterize the mineral elements, volatile components, and metabolites in wines from various regions of China. The potential chemical markers used for differentiating wines from different regions in China were identified by principal component analysis (PCA) and orthogonal partial least squares discriminant analysis (OPLS-DA). Finally, three algorithms, K-nearest neighbors (KNN), support vector machine (SVM) and random forest (RF) were applied to model both single and combined datasets to ascertain the origin of wine. This classification model has demonstrated strong potential for predicting the geographical origin of Chinese red wine, which may enhance the stability of the wine market.

2. Materials and methods

2.1. Chemicals and reagents

Hydrochloric acid solution, hypomethyl blue indicator solution, and Ferrin solution in the basic physical and chemical experiment were purchased from Shanghai Amperexperiment Technology Co. (Shanghai, China). Sodium hydroxide solution, sodium hydroxide standard solution, glucose standard solution, and phenol standard solution were purchased from Sinopharm Chemical Reagent Co. (Shanghai, China). All of these chemicals were analytical grade.

Multi-element standard solutions (Ag, Al, As, Ba, Be, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, Pb, Se, Sb, Tl, V, and Zn) and single-element standard solutions (In and Rh) were obtained from Bailiwick Technology, Ltd. (Beijing, China). Sc, In, and Bi solutions were used as internal standards. Nitric acid (HNO_3, w = 65%) was purchased from Merck (Darmstadt, Germany), and hydrogen peroxide (H_2O_2, w = 30%) was purchased from Sinopharm Chemical Reagent Co (Shanghai, China). All of these chemicals were analytical grade.

To analyze volatile compounds, 3-octanol (purity ≥97.0%, analytical grade) and sodium chloride (analytical grade) were obtained from Sigma-Aldrich (St. Louis, MO, USA) and Sinopharm Chemical Reagent Co., Ltd., respectively.

For metabolomic analysis, methanol, acetonitrile, and ammonia were obtained from Thermo Fisher Scientific (Waltham, MA, USA). Ammonium acetate, used as the mobile phase additive, was purchased from Sigma Aldrich. All of these chemicals and standards were high performance liquid chromatography (HPLC) grade. Ultrapure water was obtained using a Milli-Q water purification system (>18.2 MΩ, Millipore, Billerica, MA, USA).

2.2. Red wine samples

In this study, 90 bottles of Cabernet Sauvignon dry red wine from famous producing areas in China were collected and analyzed. All authentic wines were purchased directly from the manufacturers and were from three different production areas: Bohai Bay (BHW), the eastern foothills of the Helan Mountains in Ningxia (HLS), and the Huai Zhuo Basin (HZ). A total of thirty samples were collected from each region. Sample vintages spanning 2016 to 2021 were collected from original single-varietal wines to ensure geographical typicality, and all samples were stored in cold storage at 4 °C until analysis.

2.3. Determination of physical and chemical indexes

Physical and chemical parameters of wine samples - alcohol values, total sugar, total acid, and pH values - were determined with reference to the previous method (OIV, 1990; Machado de Castilhos, Cattelan, Conti-Silva, & Del Bianchi, 2013).

2.4. Mineral element determination

For mineral element analysis, pretreatment of wine samples was carried out according to the previous method with some modifications (Sun et al., 2023). Firstly, 15 mL of wine was placed in a 50 mL centrifuge tube, to which 5 mL of 68% HNO_3 was added. The sample was soaked overnight; the next day, the centrifuge tube was placed on a graphite ablator (SH230, Jinan Haineng Instrument co., Ltd., Shandong, China) and heated to ablution at 120 °C. After 1 h, the tube was removed and cooled, 2 mL of 30% H_2O_2 was added, and the tube was completely ablated; after 1 h, the cap was opened and the liquid was heated up to 1 mL. Ultrapure water was added to bring the volume of the sample to 50 mL, which was used as the original digestion solution.

The As, Ba, Cd, Co, Cr, Cu, Ni, Pb, Sb, Tl, V, and Zn contents of the wine digestion solution were determined using inductively coupled plasma MS (NexION 350×, PerkinElmer, Waltham, MA, USA). The instrument parameters were as follows: radiofrequency power, 1100 W; plasma gas flow rate, 15 L/min; carrier gas flow rate: 0.94 L/min; auxiliary gas flow rate: 1.2 L/min; and lens voltage: 6.0 V; sampling flow rate: 0.8 mL/min. Using 2% HNO_3 as the medium, the elemental standard solution was diluted by step by step and a standard curve was plotted (linear range: 1.0, 2.0, 5.0, 10.0, 20.0, 50.0, 100.0, 250.0, and 500.0 μg/L). Sc, In, and Bi standard solutions were used as internal standard solution (10 μg/L), added through internal standard tubes.

2.5. Volatile compound determination

Volatile compounds in wine samples were semi-quantitatively analyzed using headspace solid-phase micro-extraction and gas chromatography–mass spectrometry techniques (HS-SPME-GC–MS). Moreover, 8 mL of wine sample and 10 μL of 3-octanol were placed in a 20-mL headspace vial containing 2 g NaCl; the headspace vial was capped tightly, vortexed and oscillated for 3 s, and immediately preheated for 15 min at 40 °C in a water bath, followed by insertion of an extraction needle (50/30 μm, SUPELCO Company, USA) to adsorb volatile components for 40 min. The samples were manually injected into the apparatus to resolve the samples for 6 min at 240 °C in a gasification chamber (Wang et al., 2023).

Volatile compounds were analyzed using a Gas Chromatography-Mass Spectrometer (5977 A/7890B, Agilent Technologies, Santa Clara, CA, USA) equipped with a strong polarity column (HP-INNOWAX, 60 m × 0.25 mm × 0.25 μm, Agilent Technologies, USA). The GC conditions were as follows: the carrier gas was high-purity helium (≥99.999%), the flow rate was 2 mL/min, and the sample was injected without a shunt. The heating procedure was as follows: the starting temperature was 40 °C and was increased to 80 °C at 3 °C/min, maintained at 3 °C/min for 6 min, and then increased to 240 °C at 5 °C/min. The MS conditions

were as follows: ion source temperature of 230 °C; quadrupole temperature of 150 °C; transmission line temperature of 250 °C; electron energy of 70 eV; and mass scanning range from m/z 29 to 300.

Volatile compounds were identified based on comparison with the NIST14 spectral library. Semi-quantitative analysis was used to quantify all volatile components, and the relative content of volatile components was calculated based on the ratio of the peak area of each compound to the peak area of the internal standard (3-octanol) (Wang et al., 2023).

2.6. Untargeted metabolomics determination

For metabolite analysis, pretreatment of wine samples was carried out according to the previous method with some modifications (PAN, GU, LV, et al., 2022). Add 0.5 mL of sample to pre-cooled methanol/acetonitrile/water solution (2:2:1, *v*/v) vortex to mix, and cryogenically sonicate for 30 min. The sample was placed at −20 °C for 10 min. And a low-temperature high-speed centrifuge (5430R, Eppendorf AG, German) was used to centrifuged the sample at 140,00*g* 4 °C for 20 min. The supernatant was dried under vacuum, after which 100 μL of acetonitrile water solution (acetonitrile:water = 1:1,v/v) was added to re-dissolve the sample, followed by vortexing and centrifugation at 14,000*g* 4 °C for 15 min. The supernatant was collected for analysis.

Red wine extracts were analyzed using an Agilent 1290 Infinity HPLC system with a Waters ACQUITY UPLC BEH Amide column (100 mm × 2.1 mm, 1.7 μm; Agilent Technologies, USA). The mobile phases used were (A) water containing 25 mM ammonium acetate and 25 mM aqueous ammonia and (B) acetonitrile. The column temperature was 25 °C, flow rate was 0.3 mL/min, and injection volume was 2 μL. The gradient elution was as follows: 0–1.5 min, 98% B; 1.5–12 min, linear change from 98% to 2% B; 12–14 min, 2% B; 14–14.1 min, linear change from 2% to 98% B; 14.1–17 min, 98% B.

A Q Exactive mass spectrometer (HF-X, Thermo Fisher Scientific, USA) was used for MS analysis, and positive and negative electrospray ionization (ESI+/ESI−) modes were used for detection. The parameters of the ESI source and MS were as follows: Atomization gas auxiliary heating gas 1 (Gas1): 60, auxiliary heating gas 2 (Gas2): 60, CUR:30 psi, ion source temperature: 600 °C; and spray voltage (ion spray voltage floating) ±5500 V (positive and negative modes). The detection range of the first-stage mass-charge ratio was 80–1200 Da, resolution was 60,000, and scanning accumulation time was 100 ms. The second stage involved a segmented acquisition method with a scanning range of 70–1200 Da, a secondary resolution of 30,000, a scanning cumulative time of 50 ms, and a dynamic exclusion time of 4 s.

The raw data were converted into mzXML (Mass Spectrometry Data eXtensible Markup Language) format using ProteoWizard (http://proteowizard.sourceforge.net/). Peak alignment, retention time correction, and peak area extraction were performed using XCMS (eXtensible Computational Mass Spectrometry). Initially, the data extracted by XCMS were subjected to metabolite structure identification and data preprocessing. The quality of the experimental data was evaluated, and the data were analyzed.

2.7. Statistical analyses

The distribution trend of red wine samples from different production regions was visualized using SIMCA-P 14.1 software (Sartorius, Göttingen, Germany) using principal component analysis (PCA) and an orthogonal partial least squares discriminant analysis (OPLS-DA) model. Significant differences were analyzed using multiple comparison tests with SPSS version 23.0 software (SPSS, Inc., Chicago, IL, USA). Origin 2018 software was used to plot heat maps (OriginLab, Northampton, MA, USA).

Three machine learning algorithms, K-nearest neighbors (KNN), support vector machine (SVM), and random forest (RF), were implemented using SPSSPRO software. To objectively evaluate whether the models were useful for geographic identification, red wine samples were randomly divided into a training set (70%) and test set (30%). Model performance was evaluated in terms of various metrics, such as accuracy, precision, recall, and F1 score.

3. Results and discussion

3.1. Physicochemical index analysis

The results of physicochemical indexes of wines from different origins, including alcohol content, total sugar, total acid, and pH values, are shown in Supplementary Table S1. A one-way ANOVA analysis revealed that alcohol content in wine samples from the three origins ranged from 13.07 to 14.61%, total sugar content varied from 3.01 to 3.98 g/L, total acid content spanned from 3.73 to 5.67 g/L, and pH values were between 3.55 and 3.86. Upon examining the chemical composition of wine samples from the three appellations, it was determined that there were no significant differences in total sugar and total acid contents among Cabernet Sauvignon wines from these regions. However, significant differences were noted in alcohol content between the BHW and HLS appellations, and variations in pH were observed across all three areas. The analysis of physicochemical parameters alone proved insufficient for distinguishing wines from different origins, necessitating further analysis.

3.2. Mineral element fingerprinting

Various mineral elements were analyzed to explore the feasibility of tracing wine from different production areas. Supplementary Table S2 shows the average concentrations of the 12 mineral elements in Cabernet Sauvignon wines from BHW, HLS, and HZ. Multiple comparative analyses showed that the contents of nine elements, Ba, Cd, Co, Cu, Ni, Sb, Tl, V, and Zn, differed significantly among wines from the three production areas ($p < 0.05$), whereas the contents of three elements, As, Cr, and Pb did not ($p > 0.05$). As shown in Supplementary Table S2, the levels of Ba, Cd, Co, Cu, Ni, Sb, Tl, and V were significantly higher in BHW wines than in HZ wines. This is likely because of the higher temperatures in BHW, which result in greater evaporative losses and greater water absorption by the grapes, thus increasing the concentration of trace elements in the grapes (Greenough, Mallory-Greenough, & Fryer, 2005). Wines from HZ showed a significantly higher Zn content than did wines from other producing areas, with HLS showing the lowest Zn content. This may be because the average content of elemental Zn in the soils of HZ was higher than that of the other two areas (Guo et al., 2023; Liang et al., 2015; Zhou et al., 2022). Wine samples from different production areas had different distributions of mineral elements. Additionally, the standard deviations of several elements were large, indicating wide variations in different wine samples from the same production area. Although the analysis of variance results showed that the mineral contents of the three wine regions differed, this method could not differentiate the wine regions; further modeling is needed to discriminate the mineral contents of the wine samples.

Fig. 1A shows a heat map of the differential distribution of 12 mineral elements in the wine samples. Specifically, As, Ba, Tl, and Zn were present at high concentrations in most red wine samples, whereas the other elements were present at low concentrations. Unexpectedly, minerals other than Pb and Sb were present at high concentrations in a small percentage of BHW red wine samples. This may be related to the climate of the year in which the sample is located, or caused by a variety of stainless steel, brass, wood and plastic containers in the winery (Hopfer, Nelson, Collins, Heymann, & Ebeler, 2015).

A wine region identification model was developed using PCA and OPLS-DA to explore the applicability of mineral elements in wine region traceability. PCA, as an unsupervised identification method, can downscale complex data to provide accurate classification. Two principal components were extracted with R^2X and Q^2 values of 0.651 and 0.447, respectively, which together accounted for 65.1% and 44.7% of

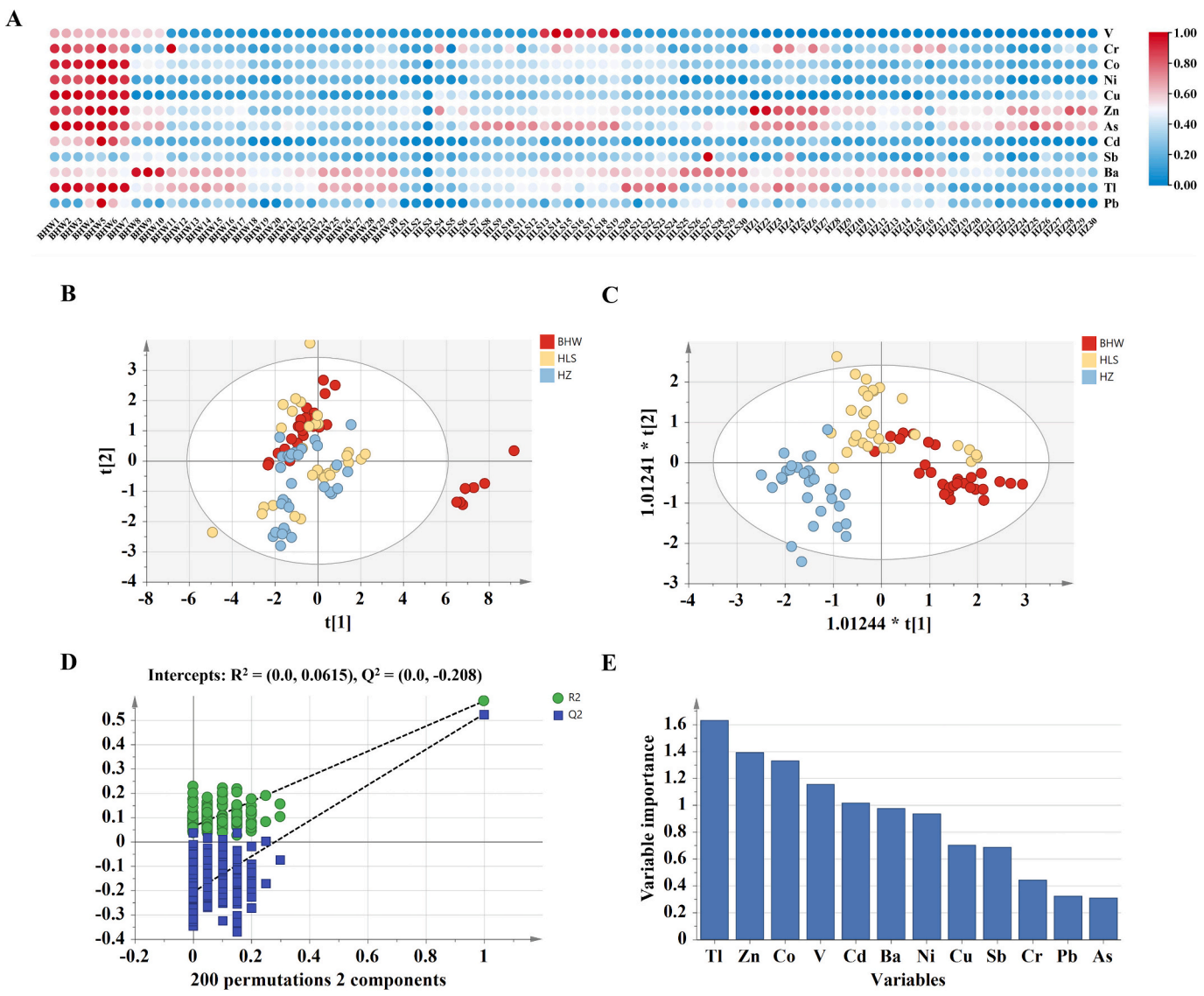


Fig. 1. Heat map (A) of the differential distribution of 12 mineral elements in wine samples (geographic sources: BHW, Bohai Bay; HLS, Ningxia Helanshan Dongluo; and HZ, Huai Zhuo Basin); score plots of the principal coordinate analysis (PCA) model (B) and the orthogonal partial least squares discriminant analysis (OPLS-DA) model (C) for the mineral elements of wines; results of cross-validation of the 200 calculations using the permutation test (D); variable importance in projection (*VIP*) plots (E).

the total explainable and predictable variance, respectively. As shown in Fig. 1B, wine samples from the BHW, HLS, and HZ appellations were difficult to differentiate, and the Q^2 parameter was <0.5, indicating that the model was poorly adapted and predictive of the geographic traceability of the wines. To overcome this issue, supervised models such as OPLS-DA have been developed to further construct the classification model.

Compared to the PCA model, a better OPLS-DA model was fitted with the main parameters $R^2X = 0.851$, $R^2Y = 0.63$, $Q^2 = 0.586$, indicating that the model had a strong interpretation and prediction ability. As shown in Fig. 1C, wine samples from different geographical sources achieved relatively good separation, and only a few samples were mixed and difficult to identify. To avoid overfitting, the OPLS-DA model was tested using 200 substitution tests. As shown in Fig. 1D, the intersection point of the Q^2 regression line and vertical axis was <0, indicating that the model was not overfitted and that model validation was effective. To screen the characteristic mineral elements with an important influence on the differences in production areas, further variable importance in projection (*VIP*) analysis was performed, and the distribution of *VIP* values of important mineral elements based on OPLS-DA was determined. A variable with a *VIP* value >1.0 was considered as a key component for classification. As shown in Fig. 1E, five variables with the highest identification potential ($VIP > 1$, $p < 0.05$) were identified (Cd, Co, Tl, V, and Zn), among which Co, Tl, and Zn ($VIP > 1.2$) were the most important discriminants. Geana et al. (2013) successfully distinguished wines from three Romanian regions using elements such as Ni, Ag, Cr, Sr, Cu, Rb, Mn, Pb, Zn, Co, and V. Similarly, Coetzee, van Jaarsveld, and Vanhaecke (2014) successfully classified wines from various South African estates using nine elements—B, Ba, Cs, Cu, Mg, Rb, Sr, Tl, and Zn—in conjunction with PCA, cluster analysis, and discriminant analysis. These findings align closely with the results of the current study. These elements are mainly influenced by local climate and soil texture. In practice, priority should be given to these important elements, so as to reduce the testing cost and improve the recognition efficiency.

3.3. *Volatile composition fingerprinting*

The volatile components of wines from the three different production

regions in China were determined using HS-SPME-GC–MS. A total of 74 volatile components was detected, including 27 alcohols, 26 esters, 4 organic acids, 4 aldehydes, 2 ketones, 3 phenols, 5 terpenes, 3 alkanes. The contents and statistics of the main volatile components are listed in Table 1. Combined with the radar plots of volatile component composition and mass concentration (Fig. 2A, B), the volatile component types and mass concentrations of wines from the three production regions in China showed some variability. Alcohols and esters were the main volatile components in the wine samples (Fig. 2A), which is consistent with the results of other studies (Garcia-Carpintero, Sanchez-Palomo, & Gonzalez-Vinas, 2011). In terms of the mass concentrations of the volatile components (Fig. 2B), HZ showed a higher content of alcohols, esters, and phenols. Different alcohols can produce wines with different aromas and flavors. Low concentrations of higher alcohols can impart a pleasant aroma to wine, whereas high concentrations of higher alcohols can have a negative impact on wine aroma and even harm human health. As the content of higher alcohols increases, botanical and black pepper aromas in wines are enhanced, whereas red fruit and woody aromas in wines are significantly suppressed (Sun & Xiao, 2018). Therefore, the content of higher alcohols, which are byproducts of wine fermentation, can be important indicators of wine quality. This study found that the alcohol richness of HZ was higher than that of wines from the other two appellations, mainly because HZ wines contained significantly higher levels of phenylethanol, furfuryl alcohol, 2,3-butanediol, and 4-terpineol compared to those in BHW and HLS. Furfuryl alcohol gives the wine a caramel aroma, which is mainly converted from furfural in the oak during aging; phenyl alcohol has a floral and peachy-fruity aroma; 4-terpineol, which is often found in the skin of grapes and vines, is warm and peppery, with lighter earthy and woodsy notes; and 2,3-butanediol, a by-product of alcoholic fermentation, has a creamy and buttery flavor, which is greater in greater amounts of sugar in the grape juice (Cadahia, Fernandez de Simon, Sanz, Poveda, & Colio, 2009). HZ showed the highest 2, 3-butanediol content, indicating that the grape berries had a high sugar content and moderate sour and sweet. This is because of the sandy soil, dry climate, abundant sunshine, and large temperature differences between day and night, which are favorable for sugar accumulation. For the same reason, the phenolic content of wines from HZ was higher than that of wines from the other two sites. Excessive volatile phenolics can diminish the fruity aroma of wines, but at low concentrations, it is generally considered that volatile phenolics $\leq$420 μg/L increase the complexity of the wine aroma (Silva, Campos, Hogg, & Couto, 2011). Esters in HZ wines with higher richness than those in other two sites belonged to the fatty acid ethyl esters (ethyl butyrate, ethyl heptanoate, ethyl isovalerate, diethyl butanedioate, monoethyl butanedioate, and ethyl 2-methylbutanoate), which are mainly produced via the fatty acid acyl and acetyl coenzyme A pathways.

To assess the volatile components in detail, PCA was performed on the volatile component data. A total of four principal components was extracted, explaining 63.5% of the total variance. Notably, the Q^2 value was −0.016. In addition, as shown in Fig. 2C, wine samples from the BHW, HLS, and HZ appellations showed high similarity and a negative Q^2 parameter, suggesting that the model was unable to provide effective geographic traceability for wine samples. To exclude intra-group differences, highlight inter-group differences, and maximize the differentiation of wines from the three appellations, supervised OPLS-DA was used, with the main parameters $R^2X = 0.518$, $R^2Y = 0.54$, and $Q^2 = 0.385$, indicating that the model fit was high but the model's predictive ability was poor. As shown in Fig. 2D, wines from the HZ were clearly separated from wines from the other two appellations, and those from BHW and HLS overlapped to some extent. The results of the 200 permutation tests were unsatisfactory (intercepts of R^2 and Q^2 were 0.149 and − 0.281, respectively; the original R^2 and Q^2 were not always larger than their corresponding values after permutation), and the OPLS-DA model showed some overfitting (Fig. 2E). In addition, to identify variables with an important contribution to sample classification, this study also calculated the *VIP* values of the volatile components based on the OPLS-DA model; the 14 compounds with *VIP* values >1 are listed in Fig. 2F. The condition of $p < 0.05$ was also considered, and 13 characteristic volatile components were screened out. Among these were 6 alcohols, 1 organic acid, 1 ester, 3 terpenes, 1 phenols, 1 aldehyde. Similar to other studies, the regional diversity in this study depends mainly on alcohols, terpenoids, ketones, and some acids and esters (Zhang et al., 2023). The results of this paper suggest that the volatile components of wines from different regions vary considerably. These differences may be influenced by the expression of functional genes, nutritional status of the grapes, and various fermentation factors (Ling et al., 2022). This study demonstrated that GC–MS-based volatile component analysis combined with multivariate statistical analysis may be used as a potential tool to identify the origin of wine.

3.4. Untargeted metabolomics fingerprinting

Untargeted metabolomics is a powerful analytical strategy for identifying the markers of food authenticity and geographic traceability. Metabolites identification is essential for obtaining information on the classification of samples and possible markers of authenticity. To further categorize wines by appellation, an untargeted metabolomics strategy based on UHPLC-Q-Exactive Orbitrap-MS was established to obtain comprehensive information on wine metabolites. Fig. 3A and B show the total ion flow chromatograms of wine samples from different geographical sources. A large number of compounds was detected in both positive and negative ion modes, indicating that the method is effective for comprehensively characterizing wine compounds. However, it is difficult to identify the geographical origin of wine samples based on macroscopic comparisons of metabolic fingerprints, both in positive and negative ion modes, although wines originate from different geographical sources. Therefore, further mining is essential to exploit the taxonomic potential of the obtained metabolic fingerprints. According to previous studies, anthocyanin glycosides and anthocyanidins in wines likely ionize in positive ion mode, whereas phenolic acids, flavonols, and flavan-3-ols produce stronger signals in negative ion mode (Palade, Croitoru, Albu, Radu, & Popa, 2021). Thus, the two polarity datasets were analyzed to obtain as many metabolites as possible.

A total of 6830 (for ESI+ mode) and 6404 (for ESI− mode) ion peaks were extracted from each wine sample using R-package XCMS software. The data were further processed, resulting in the annotation of 2075 metabolites (1276 for ESI+ mode and 799 for ESI− mode) using the available database. A PCA-based geographic classification model was developed. Specifically, the model fitted 11 principal components with principal parameters of 0.805 and 0.695 for R^2X and Q^2, respectively, in positive ion mode (Fig. 3C), whereas 12 principal components were fitted with principal parameters of 0.764 and 0.65 for R^2X and Q^2, respectively, in negative ion mode (Fig. 3D). The red wine samples from HZ clustered independently and were clearly distinguished from samples from the other producing regions, whereas samples from BHW and HLS were difficult to separate, as they showed high overlap. The aggregation of QC samples shows that the UHPLC-Q-Exactive Orbitrap-MS analysis method has good stability and reproducibility. The red wine samples from HZ were collected at a high latitude and the overall temperature was lower than that of the other two places, which led to large differences in metabolite production between red wines from this appellation and those from the other regions. Although BHW and HLS were geographically distant, there was little difference in their metabolites, which may be related to being at the same latitude.

Because PCA cannot ignore within-group errors and eliminate random errors unrelated to the purpose of the study, a supervised OPLS-DA model was constructed to identify variables that caused separation between groups. OPLS-DA modeling was used to produce score plots for the BHW, HLS, and HZ samples, which more clearly represented the metabolite differences among the three production regions. As shown in Fig. 3E and F, in positive ion mode, 49.2% (R^2X) of the variables

Table 1
Results of GC–MS analysis of aroma components of *Cabernet Sauvignon* wines from three different production areas.

No.	Compounds	CAS	Threshold (μg/L)	RI	Concentration (μg/L)		
					BHW	HLS	HZ
1	Glycerol	56–81-5	nd	nd	2358.95 ± 285.80[a]	1456.64 ± 252.98[a]	1691.84 ± 328.36[a]
2	Phenyl alcohol	60–12-8	10000[1]	1116	628.70 ± 182.77[a]	767.07 ± 128.20[a]	1310.97 ± 148.28[b]
3	Propanol	71–23-8	50000[1]	1056	29.91 ± 8.28[b]	42.49 ± 7.64[a]	51.12 ± 8.77[ab]
4	Butanol	71–36-3	150000[1]	675	13.08 ± 1.41[a]	11.30 ± 0.51[a]	10.45 ± 1.07[a]
5	2-Methyl propanol	78–83-1	40000[1]	625	109.52 ± 14.91[b]	161.61 ± 17.68[ab]	235.91 ± 8.13[a]
6	Furfuryl alcohol	98–00-0	15000[1]	851	3.56 ± 0.42[b]	nd	16.45 ± 0.87[a]
7	Benzyl alcohol	100–51-6	200000[1]	1110	19.12 ± 3.63[a]	25.00 ± 4.78[a]	28.88 ± 3.53[a]
8	N-hexanol	111–27-3	8000[1]	868	114.16 ± 14.93[a]	139.26 ± 14.61[a]	206.56 ± 13.27[a]
9	N-octanol	111–87-5	120[1]	1290	19.00 ± 4.21[b]	18.75 ± 2.93[ab]	29.28 ± 3.70[a]
10	3-Methyl-1-butanol	123–51-3	30000[1]	736	1687.99 ± 131.62[b]	2333.11 ± 121.28[ab]	3592.59 ± 144.09[a]
11	1-Nonyl alcohol	143–08-8	600[1]	1665	15.37 ± 4.11[a]	15.16 ± 0.85[b]	14.12 ± 2.27[ab]
12	3-Methylthiopropanol	505–10-2	500[1]	928	12.97 ± 4.44[b]	15.11 ± 4.26[a]	14.64 ± 2.90[c]
13	2,3-Butanediol	513–85-9	150000[1]	788	34.99 ± 4.46[a]	48.24 ± 5.36[b]	102.09 ± 6.93[b]
14	3-Methyl-1-amyl alcohol	589–35-5	7.5[3]	1324	3.40 ± 0.56[a]	4.94 ± 1.07[a]	5.89 ± 0.88[a]
15	4-Methyl-1-amyl alcohol	626–89-1	2187[3]	853	4.16 ± 0.92[a]	3.22 ± 0.42[a]	nd
16	Cis-3-hexenol	928–96-1	910[1]	935	5.58 ± 0.77[a]	9.76 ± 1.87[a]	5.96 ± 1.16[a]
17	Trans-3-hexene-1-ol	928–97-2	400[1]	852	6.54 ± 0.60[a]	8.32 ± 1.46[a]	7.69 ± 1.09[a]
18	(2*S*,3*S*)-(+)-2,3-Butanediol	19,132–06-0	nd	nd	216.25 ± 13.84[a]	nd	132.73 ± 25.12[a]
19	(*S*)-(+)-3-Methyl-1-pentanol	42,072–39-9	1000[1]	859	3.42 ± 0.44[b]	5.99 ± 0.80[ab]	6.95 ± 1.16[a]
20	2-Ethylhexanol	104–76-7	260000[3]	1034	6.27 ± 0.13	nd	nd
21	Cis-2-Penten-1-ol	1576-95-0	nd	750	2.44 ± 0.81	nd	nd
22	5-Methyl-2-hexanol	627–59-8	nd	1593	2.04 ± 0.82	nd	nd
23	2-Heptanol	543–49-7	70[1]	850	1.24 ± 0.36	nd	nd
24	Citronellol	106–22-9	100[1]	1675	20.91 ± 3.89[a]	nd	10.73 ± 2.34[b]
25	4-Terpeneol	562–74-3	110[1]	nd	5.50 ± 1.80[b]	nd	10.20 ± 1.36[a]
26	Alpha-Terpineol	98–55-5	250[3]	1191	36.53 ± 1.55	nd	nd
27	Geraniol	106–24-1	10[3]	1849	15.59 ± 1.18	nd	nd
Alcohols (27 types)							
28	Acetic acid	64–19-7	200000[1]	1190	149.10 ± 19.07[b]	224.91 ± 12.98[ab]	350.18 ± 18.26[a]
29	Caprylic acid	124–07-2	500[1]	1180	113.28 ± 7.42[a]	118.97 ± 13.77[a]	95.19 ± 18.04[a]
30	Caproic acid	142–62-1	3000[1]	990	60.84 ± 16.50[a]	73.84 ± 25.65[a]	73.34 ± 24.47[a]
31	Decanoic acid	334–48-5	15000[1]	1373	22.49 ± 9.22[a]	15.86 ± 8.25[a]	15.64 ± 7.93[b]
Organic acids (4 types)							
32	Ethyl lactate	97–64-3	128083[2]	936	87.69 ± 15.34[a]	146.59 ± 11.78[a]	123.83 ± 15.76[a]
33	Phenylethyl acetate	103–45-7	250[1]	1258	15.90 ± 8.41[a]	19.58 ± 7.88[a]	6.14 ± 1.81[a]
34	Ethyl butyrate	105–54-4	20[1]	802	33.67 ± 13.86[b]	46.72 ± 9.34[ab]	64.15 ± 7.07[a]
35	Ethyl heptanoate	106–30-9	400[1]	1334	2.07 ± 0.38[b]	2.52 ± 1.12[b]	5.04 ± 2.14[a]
36	Ethyl octanoate	106–32-1	5[1]	1196	312.28 ± 23.88[a]	325.18 ± 28.82[a]	448.25 ± 34.86[a]
37	Methyl caproate	106–70-7	nd	1205	nd	15.31 ± 1.94[a]	6.20 ± 0.82[a]
38	Ethyl isovalerate	108–64-5	7[2]	827	14.53 ± 7.01[b]	20.55 ± 7.53[b]	50.83 ± 8.86[a]
39	Ethyl decanoate	110–38-3	200[1]	1396	31.59 ± 3.36[a]	28.11 ± 8.40[a]	42.05 ± 7.15[a]
40	Methyl octanoate	111–11-5	200[1]	1390	7.53 ± 1.68[a]	6.74 ± 2.71[a]	6.00 ± 1.62[a]
41	Methyl salicylate	119–36-8	nd	1420	27.11 ± 3.96[ab]	18.19 ± 1.60[a]	32.36 ± 2.21[b]
42	Diethyl succinate	123–25-1	1200[1]	1182	470.50 ± 86.04[b]	829.46 ± 31.21[b]	1596.46 ± 47.24[a]
43	Ethyl hexanoate	123–66-0	5[1]	1000	338.69 ± 79.83[a]	448.59 ± 33.56[a]	545.82 ± 80.13[a]
44	3-Methyl-1-butanol acetate	123–92-2	30[1]	876	162.88 ± 71.04[a]	148.09 ± 39.95[a]	209.82 ± 34.75[a]
45	Ethyl acetate	141–78-6	7500[1]	894	513.70 ± 38.05[a]	751.52 ± 65.27[a]	1305.86 ± 47.62[a]
46	Hexyl acetate	142–92-7	670[1]	1011	9.55 ± 4.53[a]	11.48 ± 6.64[a]	24.13 ± 7.58[a]
47	Ethyl valerate	539–82-2	27[2]	1607	nd	nd	6.16 ± 0.86
48	Ethyl 2-furoate	614–99-3	nd	1638	3.03 ± 1.24[b]	4.09 ± 1.19[ab]	14.16 ± 2.00[a]
49	Ethyl 2-hydroxypropionate	687–47-8	50000[3]	805	110.50 ± 7.87[a]	123.34 ± 6.64[a]	341.03 ± 15.02[a]
50	Monoethyl succinate	1070-34-4	nd	2351	329.29 ± 45.08[b]	156.04 ± 7.92[b]	238.86 ± 32.86[a]
51	Ethyl 2-methyl butyrate	7452–79-1	18[1]	1003	11.79 ± 5.01[b]	18.88 ± 7.32[b]	43.85 ± 4.41[a]
52	Whisky lactone	39,212–23-2	nd	1989	nd	nd	46.49 ± 3.99
53	Ethyl phenylacetate	101–97-3	406[2]	1822	nd	nd	10.52 ± 1.56
54	Ethyl 3-hexenoate	2396-83-0	nd	1151	nd	nd	1.23 ± 0.51
55	Isoamyl lactate	19,329–89-6	nd	nd	nd	nd	14.83 ± 3.12
56	Ethyl Laurate	106–33-2	400[1]	1595	7.00 ± 0.65	nd	nd
57	Ethyl sorbate	2396-84-1	nd	1519	nd	45.82 ± 2.27	nd
Esters (26 types)							
58	3-Octanone	106–68-3	21.4[1]	830	7.73 ± 1.36[a]	5.82 ± 1.60[a]	8.52 ± 1.44[a]
59	3-Hydroxy-2-butanone	513–86-0	30000[1]	1320	3.48 ± 0.83[b]	14.99 ± 1.07[a]	17.73 ± 4.29[ab]
Ketones (2 types)							
60	Acetaldehyde	75–07-0	500[1]	645	93.54 ± 6.68	nd	nd
61	Furfural	98–01-1	14100[1]	830	24.52 ± 7.73[a]	31.64 ± 12.77[a]	31.56 ± 8.81[a]
62	Benzaldehyde	100–52-7	2000[1]	790	15.60 ± 2.11[b]	111.09 ± 11.70[a]	14.14 ± 5.02[b]
63	3-Furfural	498–60-2	nd	828	9.66 ± 2.04[a]	32.88 ± 9.22[a]	22.65 ± 11.70[a]
Aldehydes (4 types)							
64	2, 4-Di-tert-butylphenol	96–76-4	200[1]	2325	12.10 ± 4.48[b]	17.48 ± 5.56[b]	26.43 ± 9.42[a]
65	Phenol	108–95-2	nd	1978	2.64 ± 0.70[b]	1.81 ± 1.13[a]	nd
66	M-cresol	108–39-4	nd	1077	nd	0.46 ± 0.25	nd
Phenols (3 types)							
67	Styrene	100–42-5	nd	890	4.52 ± 1.89[b]	7.59 ± 4.96[b]	11.15 ± 6.01[a]

(*continued on next page*)

Table 1 (*continued*)

No.	Compounds	CAS	Threshold (μg/L)	RI	Concentration (μg/L)		
					BHW	HLS	HZ
68	Terpene oleene	586–62-9	260[3]	1083	4.69 ± 1.13[a]	nd	3.01 ± 0.78[ab]
69	Cyclooctyl tetraene	629–20-9	nd	910	5.16 ± 0.62[a]	4.60 ± 1.78[a]	5.44 ± 1.69[a]
70	Myrcene	123–35-3	1.2[3]	993	5.15 ± 2.23	nd	nd
71	D-terpene	5989-27-5	nd	1030	5.32 ± 1.69	nd	nd
Terpene (5 types)							
72	Cycloheptane	291–64-5	nd	791	nd	11.61 ± 1.45[a]	11.36 ± 2.22[a]
73	Bis(trimethylsiloxy)methylsilane	1873-88-7	nd	nd	1.06 ± 0.23	nd	nd
74	Methyltris(trimethylsiloxy)silane	17,928–28-8	nd	nd	6.09 ± 1.90	nd	nd
Alkanes (3 types)							

Odor threshold referred to literature: [1] (Chen et al., 2022). [2] (Niu, Yao, Xiao, Xiao, & Zhu, 2017). [3] (Jiang, Xi, Luo, & Zhang, 2013)

explained 91.4% (R^2Y) of differences among wines from the three production zones, with an average predictive power after cross-validation of 89.5% (Q^2); in negative ion mode, 56% (R^2X) of the variables explained 99.1% (R^2Y) of differences among wines from the three production zones, with an average predictive power after cross-validation of 97.2% (Q^2). BHW, HLS and HZ were completely separated, indicating that OPLS-DA can effectively distinguish samples.

To prevent model overfitting, the model was validated using the replacement test with 200 responses (Fig. 3G and H). The results showed that all R^2 points from left to right were lower than the original R^2 points, and all Q^2 points were lower than the original Q^2 points. In the positive ion model, the metabolites in red wine from the three production regions conformed to $R^2 = 0.066$, $Q^2 = -0.206$; in the negative ion model, $R^2 = 0.464$, $Q^2 = -0.445$. Thus, the randomized arrangement of the model produced smaller R^2 and Q^2 values than those of the original model, indicating that the modeling was effective.

Based on the criteria of $p < 0.05$ and *VIP* score > 1.5 in univariate analysis, 31 and 20 characteristic variables were screened as differential metabolites among the three wine taxa in the ESI + and ESI− models, respectively. A complete list of these metabolites is shown in Supplementary Table S3 and Table S4. The metabolites were categorized as phenylacetones and polyketides, lipids and lipid-like molecules, organic acids and their derivatives, organic oxides, organic heterocyclic compounds, benzenes, and other compounds (Fig. 4). Fig. 5A and B show heat maps of the identified differential metabolites in the red wine samples. The samples were clustered according to the appellation, and differential metabolites were classified into four categories, which initially suggested that the identified differential metabolites can be used for geographic origin identification of Chinese red wine. To understand how these metabolites behave in samples from different production regions, the relative contents of each group of compounds in red wine samples from each production region were compared, as shown in Fig. 5C–J. For the ESI+ mode data shown in Fig. 5C–F, Groups 1 and 2 were more abundant in the HLS sample and compounds in Groups 3 and 4 were most abundant in the HZ and BHW samples, respectively. For the ESI-mode data shown in Fig. 5G–J, Group 1 and 3 compounds were most abundant in the BHW sample, Group 2 compounds were least abundant in the BHW sample, and Group 4 compounds were most abundant in the HLS sample. Group 1 and 2 compounds in ESI+ mode and group 3 and 4 compounds in ESI mode were mainly amino acids and flavonoids. Flavonoids promote color stability and are responsible for the astringency of wines, affecting their taste and mouth feel (Bimpilas, Tsimogiannis, Balta-Brouma, Lymperopoulou, & Oreopoulou, 2015). Amino acids are commonly used for yeast and bacterial growth during fermentation, and amino acid levels strongly depend on the process factors, such as the addition of nitrogen (Arapitsas et al., 2020). Additionally, microbial communities on grape skin are influenced by the wine production location and climatic conditions (Wei et al., 2022). Therefore, amino acid profiles can be used as markers to distinguish wines from different regions. The distribution of the three geographic sources in China is relatively decentralized, with large differences in eco-climatic conditions. The BHW production area is located close to the Bohai Sea and experiences abundant heat and rainfall. The HLS region has a typical continental climate, with a dry climate and large temperature difference between day and night. The HZ production area is in a semi-arid zone of a middle temperate zone, with high light, moderate heat, a large temperature difference between day and night, a cool summer, and a dry climate; the soil texture is sandy and rich in minerals. Climatic conditions, including temperature, light, and rainfall, were closely related to metabolites, such as organic acids and terpenoids, in the ESI+ and ESI− models (Martinez-Luscher, Chen, Brillante, & Kurtural, 2017; Torres, Martinez-Luscher, Porte, & Kurtural, 2020; Wang et al., 2022).

3.5. Identifying red wine regions using machine learning algorithms

The PCA and OPLS-DA models based on mineral elements, volatile components, and metabolites initially discriminated the red wine regions, particularly based on metabolites. One possible explanation for the low accuracy of individual analytical methods in identifying mineral elements and volatile components is the presence of numerous variables. However, a more plausible reason is that each method alone is insufficient to obtain a comprehensive geographic profile. Therefore, some of the red wine samples overlapped and could not be clearly distinguished. In contrast to the above classifiers, machine-learning algorithms have strong capabilities for data processing and classification construction (Gromski et al., 2014). Therefore, three classical recognition algorithms (KNN, SVM, and RF) were used to verify the reliability of the models and obtain accurate classification results. In addition, multi-technology data fusion can provide complementary information on chemical features, thus improving the accuracy of geographic traceability. This study fused multi-technology data to improve the recognition rate.

These established datasets were normalized and scaled to the same matrix for KNN, SVM, and RF analyses. Table 2 and Table 3 lists the evaluation metrics (accuracy, precision, recall, and F1-score) of the three classifiers on single datasets and fusion datasets. Precision and recall are typically used to evaluate the accuracy of different algorithms. The F1-score is a comprehensive indicator of precision and recall, usually as close to 100% as possible. In this study, the recognition rate of all the training texts is above 87.0%, indicating that these models have relatively good reliability.

In Table 2, the results from all-variable modeling indicate that the prediction accuracy for volatile compounds was the lowest, ranging from 55.0% to 87.8%, whereas that for metabolites was the highest, ranging from 93.1% to 97.0%. The accuracy of mineral elements was slightly lower than that of metabolites, ranging from 71.0% to 97.0%. Compared with the results of full variable modeling, the prediction accuracy for characteristic variable modeling has improved to various degrees. Table 3 shows that the prediction accuracy for the combination of volatile components and mineral elements was the lowest, ranging from 64.3% to 97.0%, in data-level fusion. The combination of volatile components, minerals, and metabolites yielded the best performance, achieving 100.0% accuracy across all algorithms except for KNN in the validation set. Metabolites play a major role in classification and considerably enhance model prediction, whereas volatile constituents

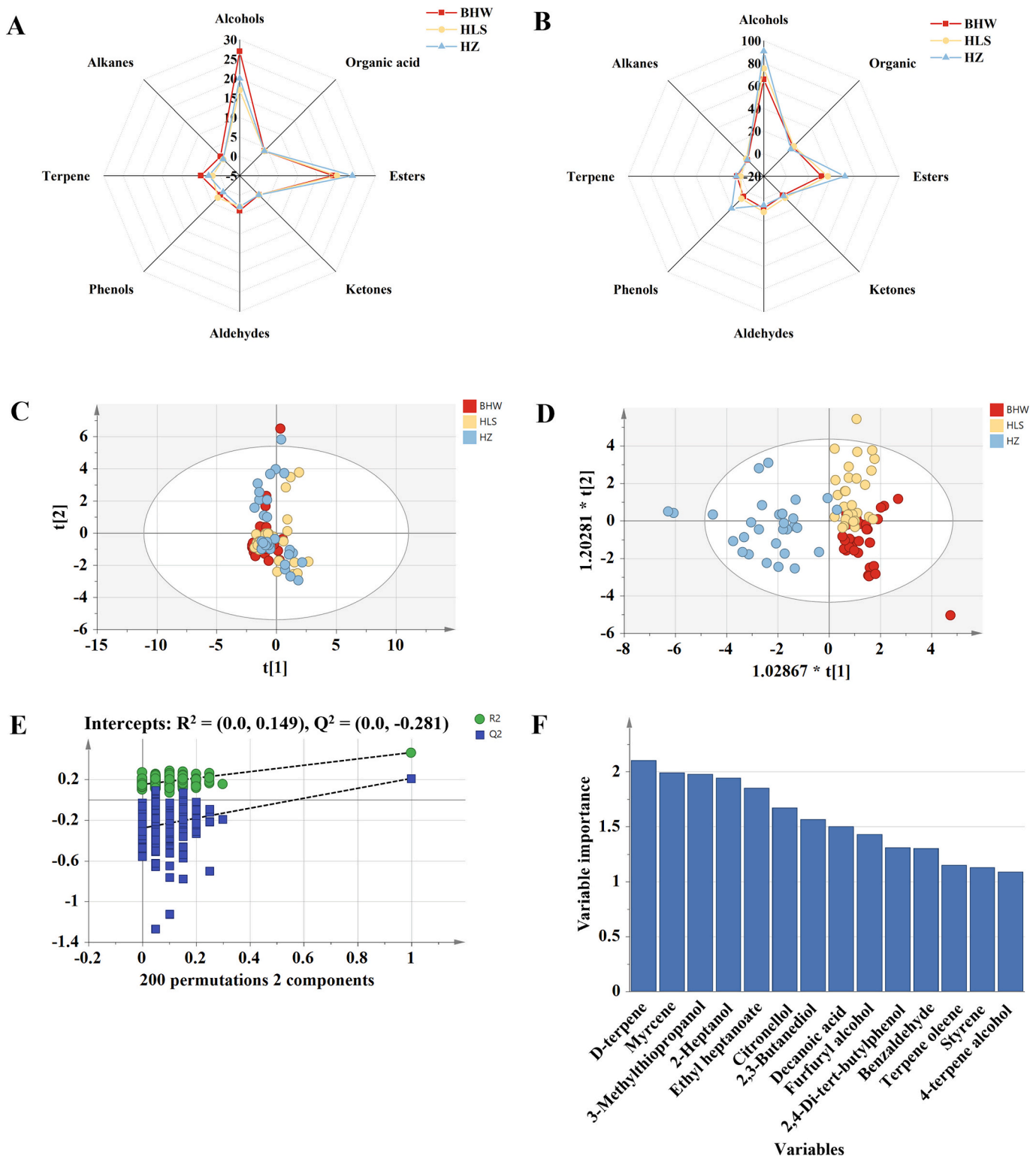


Fig. 2. Radar plots of volatile constituent species (A) and mass concentration (B) in wine samples from three production regions in China; plots of scores from the principal coordinate analysis (PCA) model (C) and the orthogonal partial least squares discriminant analysis (OPLS-DA) model (D) for volatile constituents; results of cross-validation of 200 calculations using the permutation test (E); and plots of the variable importance in projection (*VIP*) (F).

and mineral elements are comparatively less critical. In the data-level data fusion modeling process, various aspects of experimental data are effectively combined, but the fusion of data leads to an increase in the number of variables, introducing irrelevant information that can interfere with the model's discriminative ability. Consequently, this study advanced to feature-level data fusion. The results showed that datasets fused based on feature variables (mineral elements, volatile components, and metabolites) were deemed the optimal combination, achieving a 100.0% prediction rate across all three algorithms. The phenomenon may be attributed to the fact that the other three combinations have an inappropriate number of variables and contain excessive redundant information. The KNN algorithm exhibited the lowest

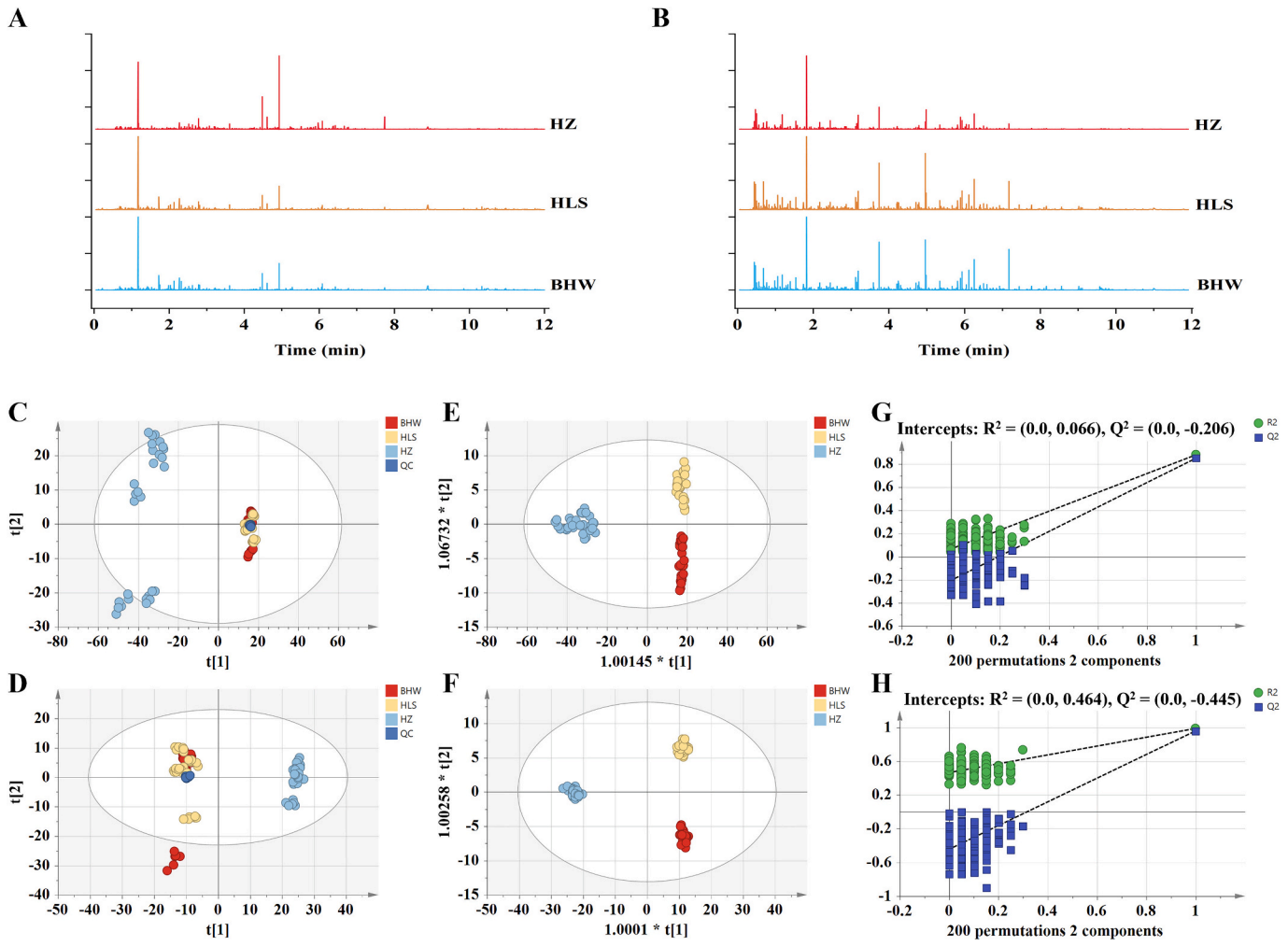


Fig. 3. Ultra-perofmance liquid chromatography-quadrupole time-of-flight-mass spectrometry (UHPLC-Q-Exactive Orbitrap MS) total ion chromatograms of Cabernet Sauvignon wines from different production areas in electrospray ionization-positive (ESI+) and ESI−negative (ESI−) modes (A and B), principal coordinate analysis (PCA) model score plots (C and D), orthogonal partial least squares discriminant analysis (OPLS-DA) model score plots (E and F), and cross-validation results of 200 calculations using the substitution test (G and H).

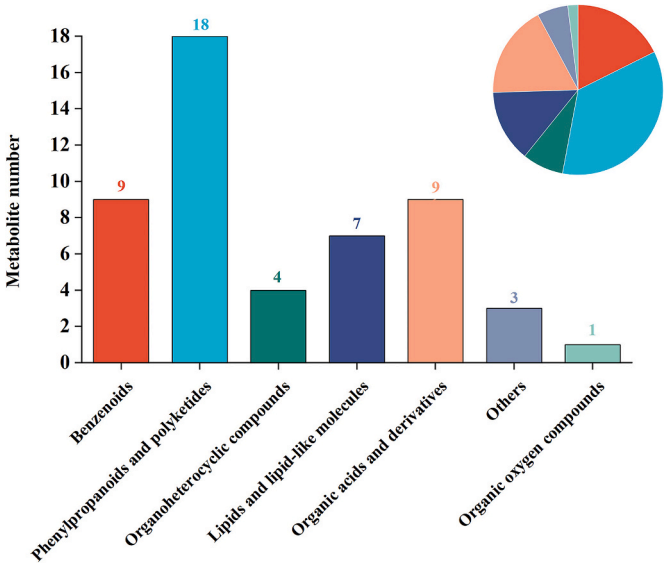


Fig. 4. Type composition of significantly different metabolites in wines from three regions.

performance, possibly because of its limitations, such as only considering the number of adjacent categories and neglecting the distance between samples. Therefore, feature-level data fusion based on mineral elements, volatile components, and metabolites is considered the best model. Similarly, Gao et al. (2022) measured the mineral elements of wines from six regions in China through ICP-MS, and combined three machine learning models (feedforward neural network, random forest and support vector machine) to identify the geographical origin of the wines, the correct recognition rates obtained using the three models were 100%, 96.67% and 98.33% respectively. The accuracy of the classification models in the above studies was lower than that obtained in this study. These results can be attributed not solely to differences among the samples but mainly to the choice of the identification technique employed. Therefore, this study indicated that ICP-MS, GC–MS and UPLC-Q-Exactive Orbitrap-MS combined with machine learning models was a potential tool to identify the geographical origin of Chinese wines.

4. Conclusions

In this study, the mineral elements, volatile components and metabolites in wine were analyzed by ICP-MS, HS-SPME-GC–MS and UHPLC-Q-Exactive Orbitrap MS analysis techniques. The most critical discriminant variables (5 mineral elements, 13 volatile components, and

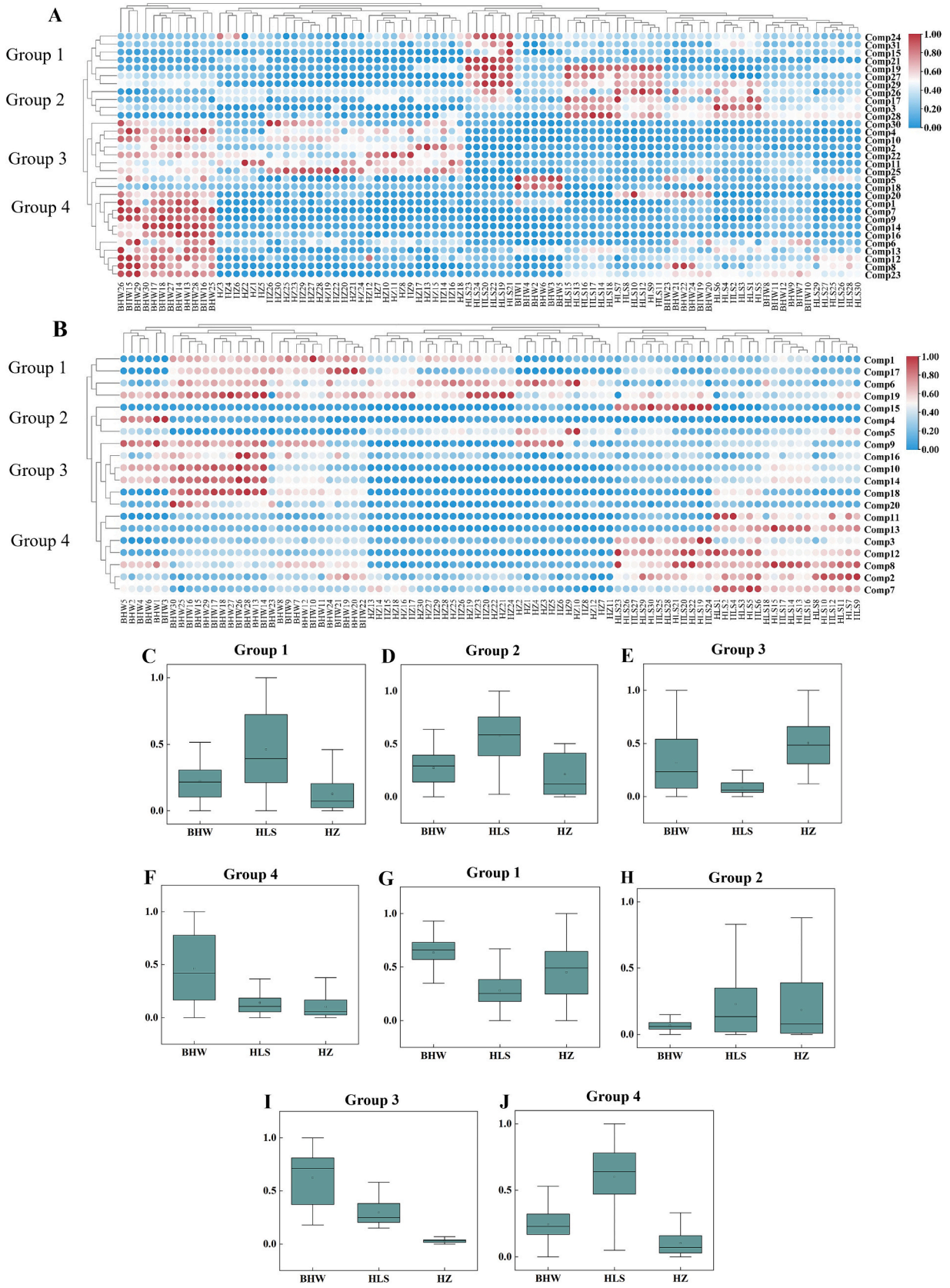


Fig. 5. Heatmap visualization of electrospray ionization-positive (ESI+) mode (A) and ESI-negative (ESI−) mode. (B) Differential metabolites in wine samples and boxplots of the relative amounts of each group of compounds in ESI+ mode (C—F) and ESI− mode (G–J) in wine samples from each region.

Table 2
Evaluation metrics of three classifiers on single datasets.

Model	Dataset	Number of variables	Evaluation indicators	Training sets			validation sets		
				KNN	SVM	RF	KNN	SVM	RF
Full-variable modeling	VC	74	Accuracy (%)	82.3	87.1	84.0	57.1	71.4	61.0
			Precision (%)	84.2	90.3	87.0	57.1	87.8	81.0
			Recall (%)	82.3	87.1	84.0	69.8	71.4	61.0
			F1-scores(%)	82.4	86.5	81.0	55.7	72.4	55.0
	ME	12	Accuracy (%)	94.0	88.7	100.0	71.0	82.1	96.0
			Precision (%)	94.0	88.8	100.0	88.0	84.4	97.0
			Recall (%)	94.0	88.7	100.0	71.0	82.1	96.0
			F1-scores(%)	93.0	88.7	100.0	72.0	82.1	97.0
	MET	2075	Accuracy (%)	100.0	100.0	100.0	93.4	96.0	93.5
			Precision (%)	100.0	100.0	100.0	93.1	97.0	94.6
			Recall (%)	100.0	100.0	100.0	94.4	96.0	93.5
			F1-scores(%)	100.0	100.0	100.0	93.7	97.0	93.6
Feature variable modeling	VC	13	Accuracy (%)	90.3	79.0	100.0	71.4	60.7	92.9
			Precision (%)	91.1	81.4	100.0	71.2	67.2	94.4
			Recall (%)	90.3	79.0	100.0	71.4	60.7	92.9
			F1-scores(%)	90.3	79.3	100.0	70.8	61.5	92.9
	ME	5	Accuracy (%)	90.3	69.4	100.0	78.6	53.6	96.4
			Precision (%)	90.3	69.7	100.0	80.4	63.7	96.9
			Recall (%)	90.3	69.4	100.0	78.6	53.6	96.4
			F1-scores(%)	90.3	67.4	100.0	77.6	53.0	96.5
	MET	51	Accuracy (%)	100.0	100.0	100.0	94.0	97.0	99.4
			Precision (%)	100.0	100.0	100.0	94.0	98.0	99.5
			Recall (%)	100.0	100.0	100.0	94.0	97.0	99.4
			F1-scores(%)	100.0	100.0	100.0	93.0	98.0	99.4

VC = Volatile compositions; ME = Mineral elements; MET = Metabolomics.

Table 3
Evaluation metrics of three classifiers on fusion datasets.

Data fusion strategy	Data set	Number of variables	Evaluation indicators	Training sets			validation sets		
				KNN	SVM	RF	KNN	SVM	RF
Data-level data fusion	VC + ME	86 (74 + 12)	Accuracy (%)	87.1	100.0	93.5	67.9	96.0	85.7
			Precision (%)	90.9	100.0	93.8	78.9	97.0	85.7
			Recall (%)	87.1	100.0	93.5	67.9	96.0	85.7
			F1-scores(%)	87.6	100.0	93.5	64.3	97.0	85.7
	VC + MET	2149 (74 + 2075)	Accuracy (%)	98.4	100.0	100.0	96.8	100.0	100.0
			Precision (%)	98.5	100.0	100.0	97.1	100.0	100.0
			Recall (%)	98.4	100.0	100.0	93.5	100.0	100.0
			F1-scores(%)	98.4	100.0	100.0	94.6	100.0	100.0
	ME + MET	2087 (12 + 2075)	Accuracy (%)	96.8	100.0	100.0	85.7	100.0	100.0
			Precision (%)	97.1	100.0	100.0	90.1	100.0	100.0
			Recall (%)	96.8	100.0	100.0	85.7	100.0	100.0
			F1-scores(%)	96.7	100.0	100.0	96.7	100.0	100.0
	VC + ME + MET	2161 (74 + 12 + 2075)	Accuracy (%)	100.0	100.0	100.0	98.9	100.0	100.0
			Precision (%)	100.0	100.0	100.0	98.8	100.0	100.0
			Recall (%)	100.0	100.0	100.0	98.9	100.0	100.0
			F1-scores(%)	100.0	100.0	100.0	98.9	100.0	100.0
feature-level data fusion	VC + ME	18 (13 + 5)	Accuracy (%)	93.5	88.7	100.0	75.0	82.1	85.7
			Precision (%)	94.6	89.1	100.0	78.7	88.5	92.9
			Recall (%)	93.5	88.7	100.0	75.0	82.1	85.7
			F1-scores(%)	93.6	88.7	100.0	75.3	80.6	86.7
	VC + MET	64 (13 + 51)	Accuracy (%)	100.0	100.0	100.0	100.0	100.0	98.7
			Precision (%)	100.0	100.0	100.0	100.0	100.0	96.4
			Recall (%)	100.0	100.0	100.0	100.0	100.0	97.8
			F1-scores(%)	100.0	100.0	100.0	100.0	100.0	97.1
	ME + MET	56 (5 + 51)	Accuracy (%)	100.0	100.0	100.0	99.4	100.0	100.0
			Precision (%)	100.0	100.0	100.0	98.5	100.0	100.0
			Recall (%)	100.0	100.0	100.0	98.4	100.0	100.0
			F1-scores(%)	100.0	100.0	100.0	98.4	100.0	100.0
	VC + ME + MET	69 (13 + 5 + 51)	Accuracy (%)	100.0	100.0	100.0	100.0	100.0	100.0
			Precision (%)	100.0	100.0	100.0	100.0	100.0	100.0
			Recall (%)	100.0	100.0	100.0	100.0	100.0	100.0
			F1-scores(%)	100.0	100.0	100.0	100.0	100.0	100.0

VC = Volatile compositions; ME = Mineral elements; MET = Metabolomics.

51 metabolites) in wine origin classification were selected by PCA and OPLS-DA. Subsequently, three algorithms were applied to model both single and fused datasets for origin identification. The results showed that the fused datasets based on feature variables, combining mineral elements, volatile components, and metabolites, achieved the best performance, with predictive accuracy of 100% across all three algorithms. This study underscores the efficacy of a data fusion strategy for authenticity identification of Chinese wine. Future research will include

more representative samples from various production areas to enhance and verify the accuracy, reliability, and practicability of the discriminant model.

CRediT authorship contribution statement

Kexiang Chen: Data curation, Writing – original draft. **Hongtu Xue:** Writing – review & editing. **Qi Shi:** Methodology. **Fan Zhang:** Resources, Software. **Qianyun Ma:** Visualization. **Jianfeng Sun:** Supervision. **Yaqiong Liu:** Software. **Yiwei Tang:** Methodology. **Wenxiu Wang:** Funding acquisition, Supervision, Validation, Writing – review & editing.

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

The authors do not have permission to share data.

Acknowledgements

This research was supported by Special Project of National Agricultural Science and Technology Park (No. 2021C-09).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochx.2024.101412.

References

Arapitsas, P., Ugliano, M., Marangon, M., Piombino, P., Rolle, L., Gerbi, V., Versari, A., & Mattivi, F. (2020). Use of Untargeted Liquid Chromatography–Mass Spectrometry Metabolome To Discriminate Italian Monovarietal Red Wines, Produced in Their Different Terroirs. *Journal of Agricultural and Food Chemistry, 68*(47), 13353–13366. https://doi.org/10.1021/acs.jafc.0c00879

Bimpilas, A., Tsimogiannis, D., Balta-Brouma, K., Lymperopoulou, T., & Oreopoulou, V. (2015). Evolution of phenolic compounds and metal content of wine during alcoholic fermentation and storage. *Food Chemistry, 178*, 164–171. https://doi.org/10.1016/j.foodchem.2015.01.090

Cadahia, E., Fernandez de Simon, B., Sanz, M., Poveda, P., & Colio, J. (2009). Chemical and chromatic characteristics of Tempranillo, cabernet sauvignon and merlot wines from DO Navarra aged in Spanish and French oak barrels. *Food Chemistry, 115*(2), 639–649. https://doi.org/10.1016/j.foodchem.2008.12.076

Cao, S. R., Du, H., Tang, B. B., Xi, C. X., & Chen, Z. Q. (2021). Non-target metabolomics based on high-resolution mass spectrometry combined with chemometric analysis for discriminating geographical origins of Rhizoma Coptidis. *Microchemical Journal, 160*, Article 105685. https://doi.org/10.1016/j.microc.2020.105685

Chen, H., Liu, Y. Q., Chen, J. W., Fu, X. F., Suo, R., Chitrakar, B., & Wang, J. (2022). Effects of spontaneous fermentation on microbial succession and its correlation with volatile compounds during fermentation of Petit Verdot wine. *LWT- Food Science and Technology, 168*, Article 113890. https://doi.org/10.1016/j.lwt.2022.113890

Coetzee, P. P., van Jaarsveld, F. P., & Vanhaecke, F. (2014). Intraregional classification of wine via ICP-MS elemental fingerprinting. *Food Chemistry, 164*, 485–492. https://doi.org/10.1016/j.foodchem.2014.05.027

Drivelos, S. A., Higgins, K., Kalivas, J. H., Haroutounian, S. A., & Georgiou, C. A. (2014). Data fusion for food authentication. Combining rare earth elements and trace metals to discriminate "Fava Santorinis" from other yellow split peas using chemometric tools. *Food Chemistry, 165*, 316–322. https://doi.org/10.1016/j.foodchem.2014.03.083

Gao, F., Hao, X., Zeng, G., Guan, L., Wu, H., Zhang, L., Wei, R., Wang, H., & Li, H. (2022). Identification of the geographical origin of Ecolly (*Vitis vinifera* L.) grapes and wines from different Chinese regions by ICP-MS coupled with chemometrics. *Journal of Food Composition and Analysis, 105*, Article 104248. https://doi.org/10.1016/j.jfca.2021.104248

Garcia-Carpintero, E. G., Sanchez-Palomo, E., & Gonzalez-Vinas, M. A. (2011). Aroma characterization of red wines from cv. Bobal grape variety grown in La Mancha region. *Food Research International, 44*(1), 61–70. https://doi.org/10.1016/j.foodres.2010.11.013

Geana, I., Iordache, A., Ionete, R., Marinescu, A., Ranca, A., & Culea, M. (2013). Geographical origin identification of Romanian wines by ICP-MS elemental analysis. *Food Chemistry, 138*(2–3), 1125–1134. https://doi.org/10.1016/j.foodchem.2012.11.104

Greenough, J. D., Mallory-Greenough, L. M., & Fryer, B. J. (2005). Geology and wine 9: Regional trace element fingerprinting of Canadian wines. *Geoscience Canada, 32*(3), 129–137.

Gromski, P. S., Correa, E., Vaughan, A. A., Wedge, D. C., Turner, M. L., & Goodacre, R. (2014). A comparison of different chemometrics approaches for the robust classification of electronic nose data. *Analytical and Bioanalytical Chemistry, 406*(29), 7581–7590. https://doi.org/10.1007/s00216-014-8216-7

Guo, G. H., Chen, S. Q., Lei, M., Wang, L. Q., Yang, J., & Qiao, P. W. (2023). Spatiotemporal distribution characteristics of potentially toxic elements in agricultural soils across China and associated health risks and driving mechanism. *Science of the Total Environment, 887*, Article 163897. https://doi.org/10.1016/j.scitotenv.2023.163897

Hopfer, H., Nelson, J., Collins, T. S., Heymann, H., & Ebeler, S. E. (2015). The combined impact of vineyard origin and processing winery on the elemental profile of red wines. *Food Chemistry, 172*, 486–496. https://doi.org/10.1016/j.foodchem.2014.09.113

Jiang, B., Xi, Z. M., Luo, M. J., & Zhang, Z. W. (2013). Comparison on aroma compounds in cabernet sauvignon and Merlot wines from four wine grape-growing regions in China. *Food Research International, 51*, 482–489. https://doi.org/10.1016/j.foodres.2013.01.001

Liang, Q., Xue, Z. J., Wang, F., Sun, Z. M., Yang, Z. X., & Liu, S. Q. (2015). Contamination and health risks from heavy metals in cultivated soil in Zhangjiakou City of Hebei Province China. *Environmental Monitoring and Assessment, 187*(12), 754. https://doi.org/10.1007/s10661-015-4955-y

Liang, Y. Z., Xie, P. S., & Chan, K. (2004). Quality control of herbal medicines. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences, 812*(1–2), 53–70. https://doi.org/10.1016/j.jchromb.2004.08.041

Ling, M., Qi, M., Li, S., Shi, Y., Pan, Q., Cheng, C., Yang, W., & Duan, C. (2022). The influence of polyphenol supplementation on ester formation during red wine alcoholic fermentation. *Food Chemistry, 377*, Article 131961. https://doi.org/10.1016/j.foodchem.2021.131961

Longobardi, F., Casiello, G., Sacco, D., Tedone, L., & Sacco, A. (2011). Characterisation of the geographical origin of Italian potatoes, based on stable isotope and volatile compound analyses. *Food Chemistry, 124*(4), 1708–1713. https://doi.org/10.1016/j.foodchem.2010.07.092

Machado de Castilhos, M. B., Cattelan, M. G., Conti-Silva, A. C., & Del Bianchi, V. L. (2013). Influence of two different vinification procedures on the physicochemical and sensory properties of Brazilian non-*Vitis vinifera* red wines. *LWT- Food Science and Technology, 54*(2), 360–366. https://doi.org/10.1016/j.lwt.2013.06.020

Majchrzak, T., Wojnowski, W., & Plotka-Wasylka, J. (2018). Classification of Polish wines by application of ultra-fast gas chromatography. *European Food Research and Technology, 244*(8), 1463–1471.

Marchionni, S., Braschi, E., Tommasini, S., Bollati, A., Cifelli, F., Mulinacci, N., … Conticelli, S. (2013). High-precision Sr-87/Sr-86 analyses in wines and their use as a geological fingerprint for tracing geographic provenance. *Journal of Agricultural and Food Chemistry, 61*(28), 6822–6831. https://doi.org/10.1021/jf4012592

Martinez-Luscher, J., Chen, C. C. L., Brillante, L., & Kurtural, S. K. (2017). Partial solar radiation exclusion with color shade nets reduces the degradation of organic acids and flavonoids of grape berry (*Vitis vinifera* L.). *Journal of Agricultural and Food Chemistry, 65*(49), 10693–10702. https://doi.org/10.1021/acs.jafc.7b04163

Mir-Cerdà, A., Granell, B., Izquierdo-Llopart, A., Sahuquillo, A., López-Sánchez, J. F., Saurina, J., & Sentellas, S. (2022). Data fusion approaches for the characterization of musts and wines based on biogenic amine and elemental composition. *Sensors, 22*(6), 2132. https://doi.org/10.3390/s22062132

Niu, Y. W., Yao, Z. M., Xiao, Q., Xiao, Z. B., & Zhu, J. C. (2017). Characterization of the key aroma compounds in different light aroma type Chinese liquors by GC-olfactometry, GC-FPD, quantitative measurements, and aroma recombination. *Food Chemistry, 233*, 204–215. https://doi.org/10.1016/j.foodchem.2017.04.103

OIV. (1990). Recueil des Methods International d'Analyse des Vins et des Mouts. Retrieved 29th September 2023, from International Organisation of Vine and Wine: http://www.oiv.int/en/statistiques.

OIV. (2022). The Statistical Information of Wine in China. Retrieved 29th September 2023, from International Organisation of Vine and Wine: http://www.oiv.int/en/statistiques.

Palade, L. M., Croitoru, C., Albu, C., Radu, G. L., & Popa, M. E. (2021). Identification of tentative traceability markers with direct implications in polyphenol fingerprinting of red wines: Application of LC-MS and Chemometrics methods. *Separations, 8*(12), 233. https://doi.org/10.3390/separations8120233

Pan, Y., Gu, H. W., Lv, Y., et al. (2022). Untargeted metabolomic analysis of Chinese red wines for geographical origin traceability by UPLC-QTOF-MS coupled with chemometrics. *Food Chemistry, 394*, Article 133473. https://doi.org/10.1016/j.foodchem.2022.133473

Plotka-Wasylka, J., Frankowski, M., Simeonov, V., Polkowska, Z., & Namiesnik, J. (2018). Determination of metals content in wine samples by inductively coupled plasma-mass spectrometry. *Molecules, 23*(11), 2886. https://doi.org/10.3390/molecules23112886

Silva, I., Campos, F. M., Hogg, T., & Couto, J. A. (2011). Factors influencing the production of volatile phenols by wine lactic acid bacteria. *International Journal of Food Microbiology, 145*(2–3), 471–475. https://doi.org/10.1016/j.ijfoodmicro.2011.01.029

Snopek, L., Mlcek, J., Sochorova, L., Baron, M., Hlavacova, I., Jurikova, T., … Sochor, J. (2018). Contribution of red wine consumption to human health protection. *Molecules, 23*(7), 1684. https://doi.org/10.3390/molecules23071684

Sun, H., Gao, P., Dong, J., Zhao, Q., Xue, P., Geng, L., Zhao, J., & Liu, W. (2023). Rhizosphere bacteria regulated arsenic bioavailability and accumulation in the soil–Chinese cabbage system. *Ecotoxicology and Environmental Safety, 249*, Article 114420. https://doi.org/10.1016/j.ecoenv.2022.114420

Sun, Z., & Xiao, D. (2018). Review in metabolic modulation of higher alcohols in top-fermenting yeast. *Advances in Applied Biotechnology, 444*, 767–773. https://doi.org/10.1007/978-981-10-4801-2_79

Torres, N., Martinez-Luscher, J., Porte, E., & Kurtural, S. K. (2020). Optimal ranges and thresholds of grape berry solar radiation for flavonoid biosynthesis in warm climates. *Frontiers in Plant Science, 11*, 931. https://doi.org/10.3389/fpls.2020.00931

Wang, X., Chen, J., Ge, X., Fu, X., Dang, C., Wang, J., & Liu, Y. (2023). Sequential fermentation with indigenous non-Saccharomyces yeasts and *Saccharomyces cerevisiae* for flavor and quality enhancement of Longyan dry white wine. *Food Bioscience, 55*, Article 102952. https://doi.org/10.1016/j.fbio.2023.102952

Wang, Z. X., Yin, H. N., Yang, N., Cao, J. H., Wang, J. K., Wang, X. F., & Xi, Z. M. (2022). Effect of vineyard row orientation on microclimate, phenolic compounds, individual anthocyanins, and free volatile compounds of Cabernet Sauvignon (*Vitis vinifera* L.) in a high-altitude arid valley. *European Food Research and Technology, 248*(5), 1365–1378. https://doi.org/10.1007/s00217-022-03961-9

Wei, R., Ding, Y., Gao, F., Zhang, L., Wang, L., Li, H., & Wang, H. (2022). Community succession of the grape epidermis microbes of cabernet sauvignon (*Vitis vinifera* L.) from different regions in China during fruit development. *International Journal of Food Microbiology, 362*, Article 109475. https://doi.org/10.1016/j.ijfoodmicro.2021.109475

Zhang, L., Liu, Q., Li, Y., Liu, S., Tu, Q., & Yuan, C. (2023). Characterization of wine volatile compounds from different regions and varieties by HS-SPME/GC-MS coupled with chemometrics. *Current Research in Food Science, 6*, Article 100418. https://doi.org/10.1016/j.crfs.2022.100418

Zhou, T. N., Wang, Y., Qin, J. Q., Zhao, S. Y., Cao, D. Y., Zhu, M. L., & Jiang, Y. X. (2022). Potential risk, spatial distribution, and soil identification of potentially toxic elements in *Lycium barbarum* L. (Wolfberry) fruits and soil system in Ningxia, China. *International Journal of Environmental Research and Public Health, 19*(23), Article 16186. https://doi.org/10.3390/ijerph192316186