TYPE Original Research
PUBLISHED 20 May 2025
DOI 10.3389/fnut.2025.1598875

Characterization and feature selection of volatile metabolites in Yangxian pigmented rice varieties through GC-MS and machine learning algorithms

Kaiqi Cheng[1], Ruonan Dong[1], Fei Pan[2], Wen Su[1], Lingjie Xi[1], Meng Zhang[1], Jingzhang Geng[1], Ruichang Gao[1,3], Wengang Jin[1]* and A. M. Abd El-Aty[4,5]*

[1]Qinba State Key Laboratory of Biological Resources and Ecological Environment, QinLing-Bashan Moun-tains Bioresources Comprehensive Development 2011 C. I. C, Shaanxi Province Key Laboratory of Bio-Resources, College of Bioscience and Bioengineering Shaanxi University of Technology, Hanzhong, China, [2]Institute of Apicultural Research, Chinese Academy of Agricultural Sciences, Beijing, China, [3]College of Food and Biotechnology, Jiangsu University, Zhenjiang, China, [4]Department of Pharmacology, Faculty of Veterinary Medicine, Cairo University, Giza, Egypt, [5]Department of Medical Pharmacology, Medical Faculty, Ataturk University, Erzurum, Türkiye

EDITED BY
Sui Kiat Chang,
Universiti Tunku Abdul Rahman, Malaysia

REVIEWED BY
Yu Xiao,
Hunan Agricultural University, China
Bailin Li,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Wengang Jin
✉ jinwengang@nwafu.edu.cn
A. M. Abd El-Aty
✉ abdelaty44@hotmail.com

RECEIVED 24 March 2025
ACCEPTED 22 April 2025
PUBLISHED 20 May 2025

CITATION
Cheng K, Dong R, Pan F, Su W, Xi L, Zhang M, Geng J, Gao R, Jin W and Abd El-Aty AM (2025) Characterization and feature selection of volatile metabolites in Yangxian pigmented rice varieties through GC-MS and machine learning algorithms. *Front. Nutr.* 12:1598875. doi: 10.3389/fnut.2025.1598875

Introduction: Pigmented rice is fascinated by consumers for its abundant phytochemicals and unique aroma.

Methods: In this study, GC–MS-based metabolomics of Yangxian colored rice varieties were performed to characterize their volatile metabolites through multivariate statistics and machine learning algorithms.

Results: Results showed that a total of 357 volatile metabolites were detected and segmented into 9 groups, including 96 organooxygen compounds (26.89%), 52 carboxylic acids and derivatives (14.57%), 42 fatty acyls (11.76%), 16 benzene and substituted derivatives (4.48%), and 11 hydroxy acids and derivatives (3.08%). Multivariate statistics screened 127 differentially abundant metabolites via PLS-DA. Principal component analysis revealed that the percentages of PC1 and PC2 were 52.48% and 27.09%, respectively. Based on differential metabolites with great multicollinearity above 0.8 and the chi-square test (20% feature numbers), only 7 metabolites were found to represent the overall metabolites among the several colored rice varieties. Four machine learning models were further used for the classification of various colored rice varieties, and random forest model was the optimum for predicting classification, with an accuracy of 0.97. Moreover, Shapley additive explanations analysis revealed that the 7 metabolites can be used as potential markers for representing the metabolomic profiles.

Conclusions: These results implied that GC–MS-based metabolomics combined with random forest might be effective for extracting key features among different pigmented rice varieties.

KEYWORDS
pigmented rice, metabolites, multivariate statistics, machine learning, volatiles

1 Introduction

Pigmented rice is a popular and healthy cereal fascinated by consumers because of the abundance of phytochemicals, which promote favorable health benefits (1). The primary foundation of skin coloration in rice is anthocyanins, which are flavonoids (2). In recent years, pigmented rice has garnered increasing attention for its abundance of bioactive compounds, particularly anthocyanins and flavonoids. These phytochemicals are known for their potent antioxidant and anti-inflammatory properties. A recent study by Callcott et al. (3) demonstrated that acute consumption of purple and red rice significantly improved plasma antioxidant capacity and reduced levels of proinflammatory cytokines in obese individuals. In parallel, global market trends have shown a steady rise in consumer demand for functional rice products (4). In addition, volatile flavors are also highly important for the assessment of rice quality because they can be used to determine the quality grade and price of rice in retail markets (5). Yangxian County (Hanzhong, China) is renowned for its native-colored rice provisions. Yangxian pigmented rice is mainly black, yellow, purple, red, or green in color and contains abundant fibers and bioactive components (6, 7). Therefore, pigmented rice has great potential for use in many applications and needs extra attention.

Currently, metabolomics has become an important topic of systems biology after genomics, transcriptomics, and proteomics (8). Metabolomics could provide complete information for observing the variations in low-molecular-weight chemicals in samples, and is targeted at revealing the relative changes among physiopathological variations, and processing circumstances (9). GC-MS is one of the most widely utilized approaches in metabolomic analysis, because of its low cost, repeatability, and simple data statistics, which are different from those of LC-MS or NMR (10, 11).

Although volatile organic compounds detected by GC-IMS and GC-MS methods can be used to discriminate the volatile compounds of various colored rice varieties, they are not sensitive enough to measure smaller variations in the same rice (12). To overcome this disadvantage, volatile metabolites have been a concern during the past decade. TianXin et al. (13) performed a metabolomic assay on rice from several different growing locations through GC-MS. Zhang et al. (14) compared the distinctive abundant metabolites among colored rice and white rice via a broadly targeted metabolomics method. Wang et al. (15) also revealed the traits of distinct odor chemicals in oats through broadly targeted GC–MS metabolomic techniques. These GC-MS-based metabolomic studies can acquire useful information and large amounts of data, which are often processed through multivariable statistics and contain redundant information. With the rapid development of artificial intelligence, machine learning technologies (decision trees, longistic regression, competitive adaptive reweighed sampling, etc.) have been applied during metabolomic studies; these methods can overcome these shortcomings to a great extent and exhibit great application prospects. For example, using machine learning, Zheng et al. (16) demonstrated the valid identification of more informative features from sophisticated metabolomic data on metabolites of honey and sugar in diets fed to mice. Metabolomics coupled with machine learning procedures for classifying the production regions or geographic origins of tea samples has also been published (17, 18).

Our previous studies explored the odor substances of five different colored rice varieties (raw, cooked, and puffed) (6, 9) and performed anthocyanin quantitative analysis, and health-promoting functions in Yangxian (19). However, the results of the metabolomic profiling of five different pigmented rice varieties are still unknown. In this study, differentially abundant metabolites in several Yangxian colored rice varieties were first investigated via GC-MS-based metabonomics combined with multivariate statistical analysis. Moreover, four machine learning approaches were also employed to extract the key significant features of the differentially abundant metabolites in different pigmented rice varieties, and their robustness for the discrimination of pigmented rice was also evaluated, with the hope of shedding additional light on the quality characteristics of Yangxian pigmented rice.

2 Materials and methods

2.1 Materials

The five colored rice varieties were grown and collected in Yangxian, which was provided by Shuangya Zhoudahei Organic Food Co., Ltd (Hanzhong, China). The varieties used were Shuangya Black (moisture content of 11.77 ± 0.08 g/100 g, starch content of 73.04 ± 0.18 g/100 g, crude protein content of 10.68 ± 0.14 g/100 g, lipid content of 2.96 ± 0.05 g/100 g, ash content of 1.56 ± 0.05 g/100 g), Shuangya Green (moisture content of 12.45 ± 0.11 g/100 g, starch content of 73.15 ± 0.19 g/100 g, crude protein content of 9.54 ± 0.11 g/100 g, lipid content of 3.49 ± 0.02 g/100 g, ash content of 1.37 ± 0.03 g/100 g), Shuangya Purple (moisture content of 11.95 ± 0.11 g/100 g, starch content of 74.82 ± 0.18 g/100 g, crude protein content of 9.12 ± 0.12 g/100 g, lipid content of 2.76 ± 0.08 g/100 g, ash content of 1.35 ± 0.06 g/100 g), Shuangya Red (moisture content of 11.57 ± 0.08 g/100 g, starch content of 74.99 ± 0.10 g/100 g, crude protein content of 9.76 ± 0.09 g/100 g, lipid content of 2.44 ± 0.03 g/100 g, ash content of 1.25 ± 0.05 g/100 g), and Shuangya Yellow (moisture content of 11.55 ± 0.08 g/100 g, starch content of 73.52 ± 0.10 g/100 g, crude protein content of 10.23 ± 0.06 g/100 g, lipid content of 3.11 ± 0.03 g/100 g, ash content of 1.58 ± 0.05 g/100 g) (20). The appearance photo of these colored rice was shown in Figure 1A.

2.2 Rice pretreatment and extraction

The pigmented rice was processed and extracted according to a modified procedure from Wang et al. (15). Thirty milligrams of tissue sample was accurately weighed into a 1.5 mL vial, and 20 μL of inner reference solution (0.3 mg/mL L-2-chlorophenylalanine in methanol) and 600 μL of methanol-water (4:1, v/v) solution were added. Two tiny steel balls were added, placed in a −80°C environment for 2 min, ground in a grinder, mixed with 120 μL of chloroform, vortexed for 2 min, placed in an ice water bath, subjected to ultrasonication at 40 kHz for 10 min, and left at −20°C for 30 min. After centrifugation at 4°C and 13,000 r/min for 10 min, 100 μL of the upper liquid was pooled and placed into a derived bottle. Quality control (QC) was performed by blending the same quantity of the sample extract solutions. The QC volume was consistent with that of the sample. A centrifugal concentration desiccator was used to evaporate the sample to dryness, and the sample was transferred to a glass derivatization vial. Then, the oxime reaction was performed. The sample was removed, 50 μL of BSTFA (1% trimethyl chlorosilane) original chemical mixture, 20 μL of n-hexane, and 10 μL of inside references (11 fatty acid methyl esters in chloroform) were added, and the mixture was vortexed. The sample was subsequently removed and kept at 25°C for 30 min for subsequent GC–MS metabolomic measurements.

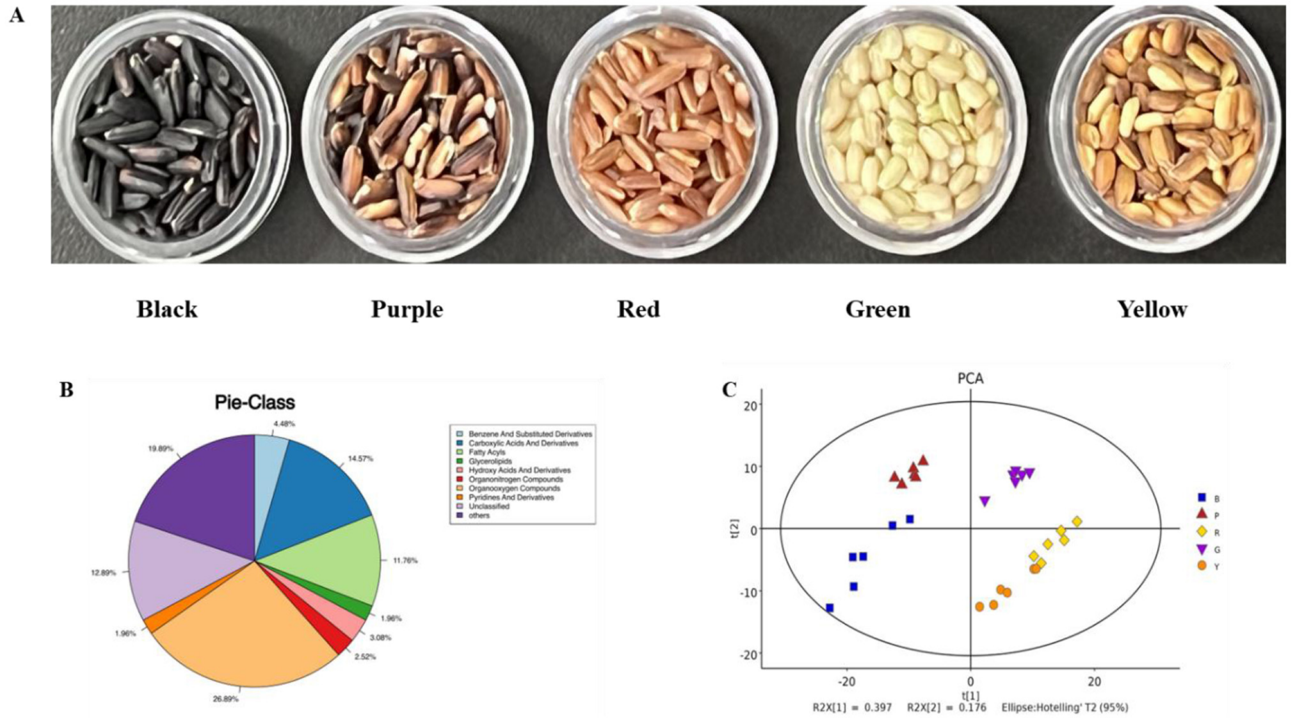

FIGURE 1
Volatile metabolomic profiles of different pigmented rice varieties. Photos of different pigmented rice varieties used **(A)**, pie-class plot of overall metabolites in different pigmented rice varieties **(B)**, and PCA score plot of overall metabolites in different pigmented rice varieties **(C)**.

Different pigmented rice samples (6 biological replicates each) were labeled according to their outer pigment, namely, black (B), purple (P), red (R), green (G), or yellow (Y).

2.3 GC–MS metabolomic assay

The chromatographic equipment used was a DB-5MS capillary column (30 m × 0.25 mm, 0.25 μm, Agilent Technologies, Inc., USA); the transport gas was helium (purity ≥99.99%), the flow rate was 1.2 mL/min; 300°C; 1 μL; the solvent delay was 5 min; the programmed temperature increase was 60°C, the temperature was maintained for 0.5 min; the temperature was elevated to 125°C at 8°C/min, and the temperature was held for 5 min; the temperature was elevated to 210°C at 5°C/min and held for 5 min; the transmission line temperature was elevated to 270°C at 10°C/min and held for 5 min; the temperature was elevated to 305°C at 20°C/min and held for 5 min; the mass spectrometry conditions: the electron ionization source was set at 330°C; and the transmission line temperature was 280°C (21).

2.4 Identification and enrichment assay of differentially abundant metabolites

To analyze the distinctness of various sample groups, a PLS-DA model with VIP distribution by SIMCA 14.1 was used. To avoid overfitting, a displacement test using 200 permutations was also carried out. The sieved differentially abundant metabolites (VIP ≥ 1, $P < 0.05$) were identified through the online KEGG database.

2.5 Machine learning approaches and simulation assessment

Decreasing the number of features hinders dimensional issues, improves model generalization and decreases overfitting. Briefly, the correlation coefficients of differentially abundant metabolites (features) were evaluated, and metabolites with great multicollinearity (above 0.8) were deleted (diminishing the features from 127 to 37). Furthermore, 20% of the features were processed via the chi-square test (diminishing the features from 37 to 7) (16). Four machine learning algorithms (XG Boost, random forest, decision tree, and longistic regression) were employed to characterize the key metabolomic profiles of different pigmented rice varieties. XGBoost was selected for its high accuracy and ability to capture complex feature interactions, whereas logistic regression offers a simple and interpretable baseline. Random forest provides robustness against overfitting in high-dimensional data, and decision trees offer a clear, interpretable classification structure (22). These algorithms were executed via the Python package (https://anaconda.org/anaconda/conda). Moreover, StratifiedKFold was carried out on the basis of the ratio of positive and negative samples, and 3-fold cross-validation was implemented, through the optimum models for observing the key metabolites among different pigmented rice varieties. After training each model, metrics such as precision, recall, accuracy, and the f1 score were calculated to evaluate the robustness of the models, as illustrated by the following equations:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\begin{array}{c}\text{True Positives} + \text{True Negatives} + \\ \text{False Positives} + \text{False Negatives}\end{array}}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.6 Shapley additive explanations (SHAP) analysis

The SHAP analysis is a "model interpretation" procedure derived from Python that explains the output of machine learning algorithms. In SHAP, an extra interpretation model is built, and all features are deemed "contributors". For every forecast, the model yields a forecast value, which can show the effect of the key features on different pigmented rice varieties and demonstrate the impact outcome (23).

3 Results and discussion

3.1 Outlook of metabolites in different colored rice varieties

To demonstrate the overall metabolomic profiles, the QC spectra and detailed metabolites in different pigmented rice samples obtained via GC-MS analysis are shown in Supplementary Figure S1. The instrumental conditions and spectra confirmed the reliability of the present results from the QC and pigmented rice samples (Supplementary Figure S1). A total of 357 metabolites were measured and segmented into 9 groups according to their chemical traits, including 96 organooxygen compounds (26.89%), 52 carboxylic acids and derivatives (14.57%), 42 fatty acyls (11.76%), 16 benzene and substituted derivatives (4.48%), and 11 hydroxy acids and derivatives (3.08%) (Figure 1B). These findings implied that organooxygen substances, carboxylic acids and derivatives, and fatty acyls were the dominant metabolites in the five pigmented rice varieties, which likely determined their unique aroma profiles.

To depict the metabolic variations identified in the five colored rice varieties, the 357 metabolites detected were analyzed via PCA. The PC1 and PC2 accounted for 39.7 and 17.6%, respectively, of the variance among these colored rice samples (Figure 1C). There was a slight overlap between red-colored rice and yellow-colored rice, and differently colored rice still relatively clustered into different groups. Moreover, hierarchical cluster analysis of the 357 metabolites was also performed, indicating that the metabolites of various colored rice varieties varied to some extent (Supplementary Figure S2). Several studies have shown that variations in metabolic patterns may be associated with genetics, varieties, nutrients, and geographical differences (13, 15). Overall, these data confirmed the high degree of similarity and high dependability among the duplicates. The distinct metabolic profiles of the five pigmented rice varieties were likely due to their genotypes and varieties.

3.2 Differentially abundant metabolites in pigmented rice

PLS-DA is frequently utilized to construct interplay equations between variables and sample categories. In the present model, all metabolites identified in several colored rice varieties were fitted via simulation, with R^2X (cum) = 0.781, R^2Y (cum) = 0.983, and Q^2 (cum) = 0.95. The majority of pigmented rice samples with various colors can be favorably classified on the PLS-DA illustration (Figure 2A), and the discrimination pattern was consistent with the PCA profile (Figure 1C). To elude over-fitting, the dependability of the PLS-DA model was validated by a permutation test, as demonstrated in Figure 2B. After 200 cross-validations, the restoration curve of simulation Q^2 crosses the abscissa, and the intercept is negative (−2.2). In all the permutation tests, R^2 and Q^2 are lower than the original values, implying that the imitation is not overfit (24). The effect of each variable for classification was further evaluated through VIP of the PLS-DA model. On the basis of VIP > 1.0 and a significance level of $p < 0.05$, a total of 127 distinctly abundant chemicals were sieved out among five different pigmented rice samples, which were further subjected to PCA and heatmap clustering analysis. The majority of the variation in pigmented rice might be distinguished via PCA, with a cumulative percentage of 79.57% (the two components were 52.48 and 27.09%, respectively) (Supplementary Figure S3). The heatmap clustering results of the top 50 metabolites in the different pigmented rice varieties are illustrated in Figure 3. As shown, the differentially abundant metabolites in black-colored rice and purple-colored rice were more similar, as both contained higher contents of protocatechuic acid, dihydroxycarbazepine, nicotianamine, uracil, etc. Similarly, black-colored rice had relatively higher contents of levan, palatinitol, maltitol, etc., than purple-colored rice. The characteristic metabolites in green-colored rice are caproic acid, pinitol, 1,7-heptanediol, etc., which are significantly distinct from those in other pigmented rice. Red- and yellow-colored rice shared the majority of the differentially abundant metabolites, but yellow-colored rice contained more metabolites, such as formoterol, 2-hydroxyadenine, and 6-aminonicotinamide (Figure 3). Ch et al. (25) used an HS-GC–MS approach for investigating metabolomic substances to classify rice samples from several states. The present results indicate that the 127 differentially abundant metabolites can also be used for classifying the overall metabolic profiles of various colored rice varieties.

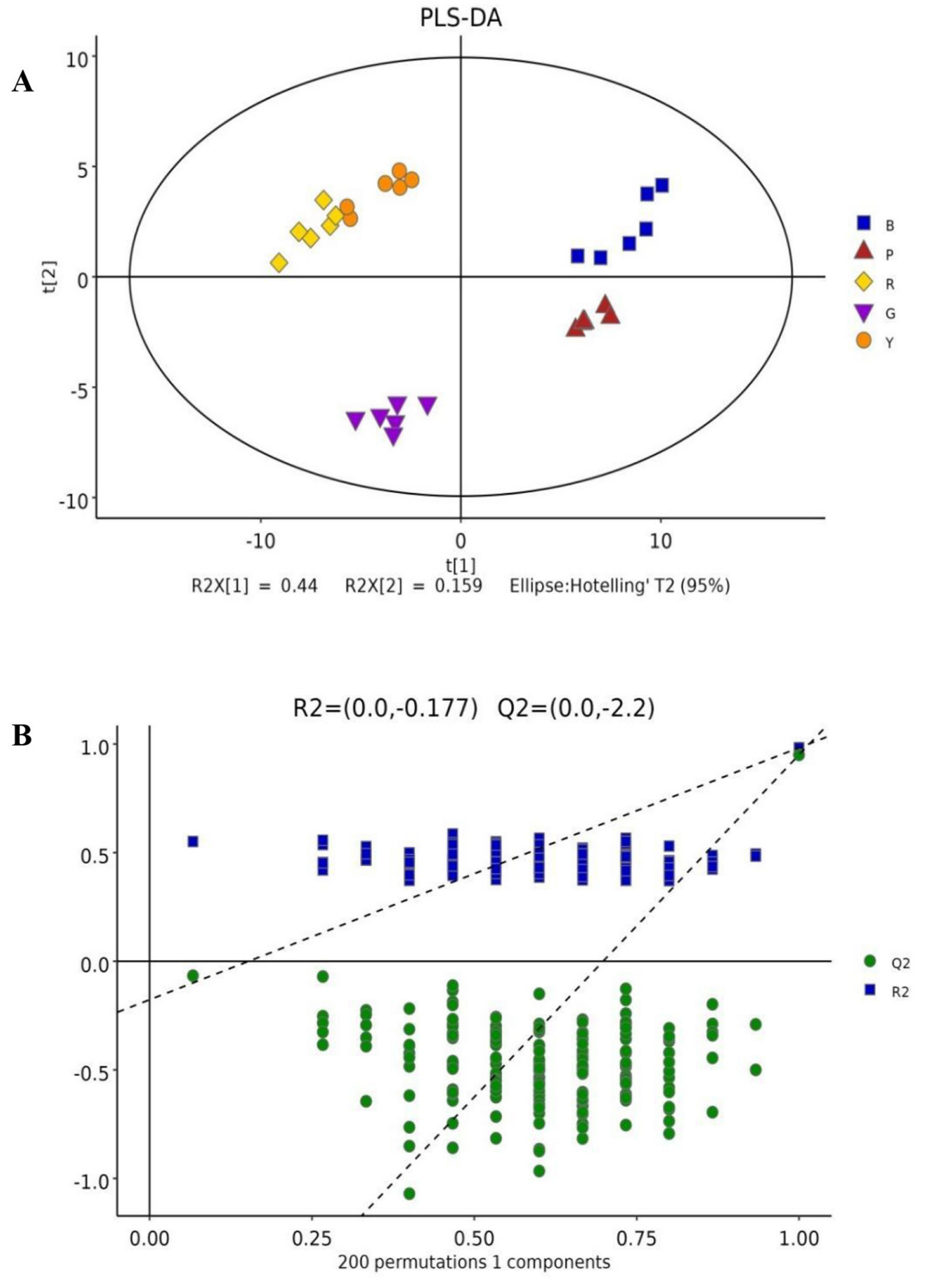

FIGURE 2
PLS-DA score plot **(A)** and cross-validation by a permutation test **(B)** of the overall metabolites in different pigmented rice varieties.

3.3 KEGG and enrichment assays of distinctly abundant metabolites

The distinct variations in metabolites are closely connected to the metabolic pathways in which they are located. The metabolic pathways that contributed the most to the differentially abundant metabolites among the different pigmented rice varieties were investigated. The KEGG database was utilized to discuss the metabolic pathway accumulation of the distinctly abundant chemicals identified in the present investigation (Figure 4). As shown, the top 3 pathways with significant differences ($p < 0.05$) were aminoacyl-tRNA biosynthesis, butanoate metabolism, and alanine, aspartate and glutamate metabolism in the different pigmented rice varieties.

Aminoacyl-tRNA biosynthesis is chiefly triggered by the direct association of an amino acid with the leading tRNA by a synthetase. Its function is to accurately target amino acids with tRNAs (26). Butanoate metabolism is a vital metabolic pathway under mild salinity stress (27). Alanine, aspartate and glutamate metabolism are in charge of osmotic balance regulation in plants. In addition, glutamate and glutamine produced by these metabolic pathways are imperative factors of phytochelatin (28). Liu et al. (29) compared pathways of differentially abundant metabolites between the normal and colored rice samples and detected alanine,

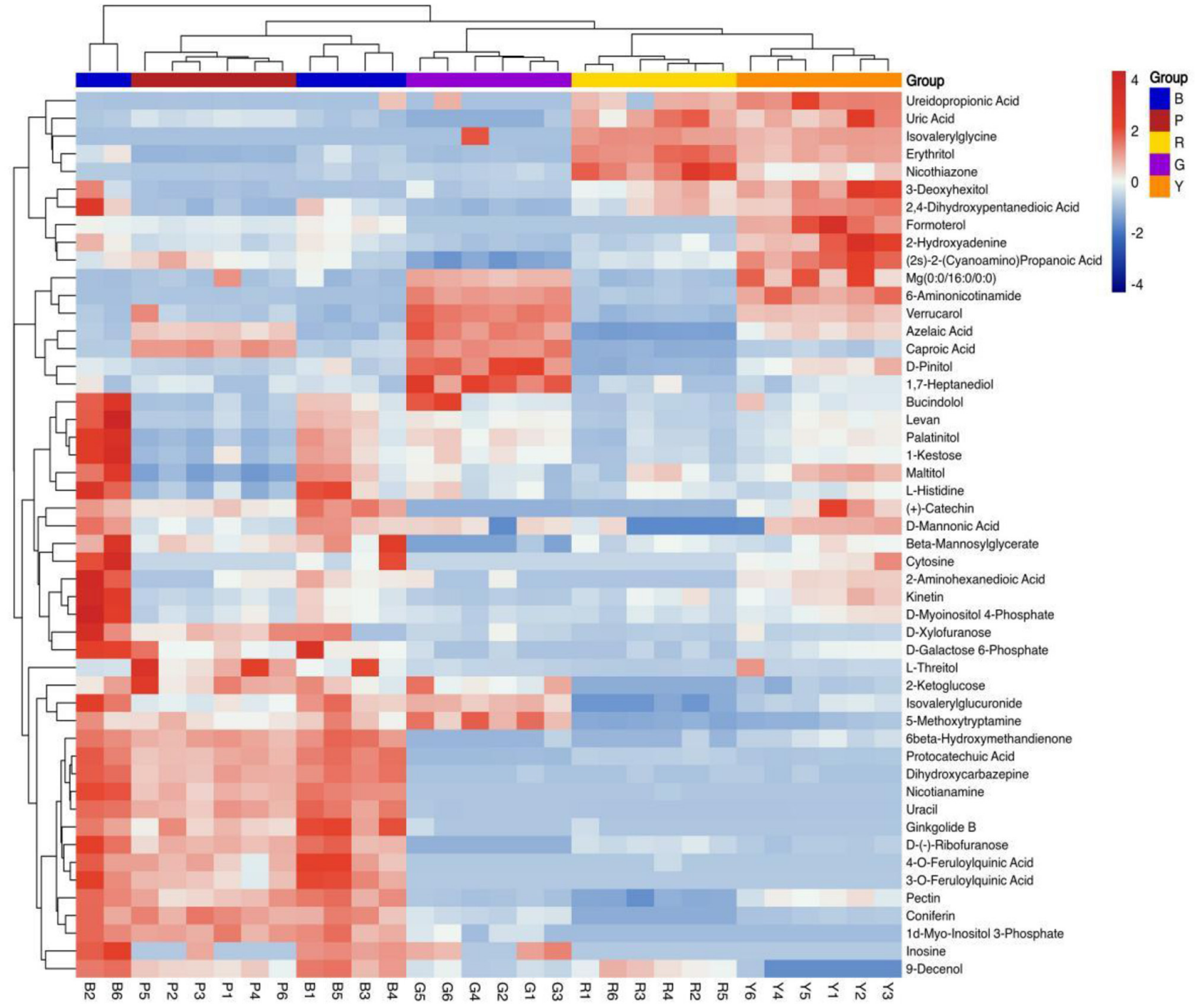

FIGURE 3
Clustering heatmap of the top 50 differentially abundant metabolites in different pigmented rice varieties.

aspartate and glutamate metabolism. Sew et al. (30) reported that pathways associated with the biosynthesis of aminoacyl-tRNA synthesis were significantly enriched in black, red, and white rice groups. The present data were almost similar to those of the previous publications.

3.4 Prediction of differentially abundant metabolites via machine learning

As 127 differentially abundant metabolites were much more abundant than the number of different pigmented rice samples (16), they still contained many redundant features, and there was still great potential for overfitting the data, leading to incorrect forecasts (16). Therefore, we processed the data, the correlation coefficients of differentially abundant metabolites (127 features) were evaluated, and metabolites with great multicollinearity (above 0.8) were deleted (diminishing the features from 127 to 37). Moreover, 20% of the features were processed through the chi-square test (diminishing the features from 37 to 7) (Supplementary Figure S4). Similar feature extraction procedures and methods were also published previously (16, 23).

Four machine learning approaches (XG Boost, random forest, decision tree, and logistic regression) were chosen to mine the data in the present work. XGBoost and random forest, as ensemble methods, are known for their high accuracy and robustness. The decision tree offers simplicity and interpretability, whereas logistic regression provides a solid baseline for binary classification tasks (23). Figures 5A–D shows the four selected machine learning models for predicting the classification of different pigmented rice varieties, and they all show better classification effects with the help of the 7 feature metabolites. Compared with the other models, the random forest model was the best (Figure 5B). Furthermore, the metrics of the four machine learning approaches via the three-fold cross-validation procedure were also evaluated, and the results are shown in Table 1. The outcomes for four assessment

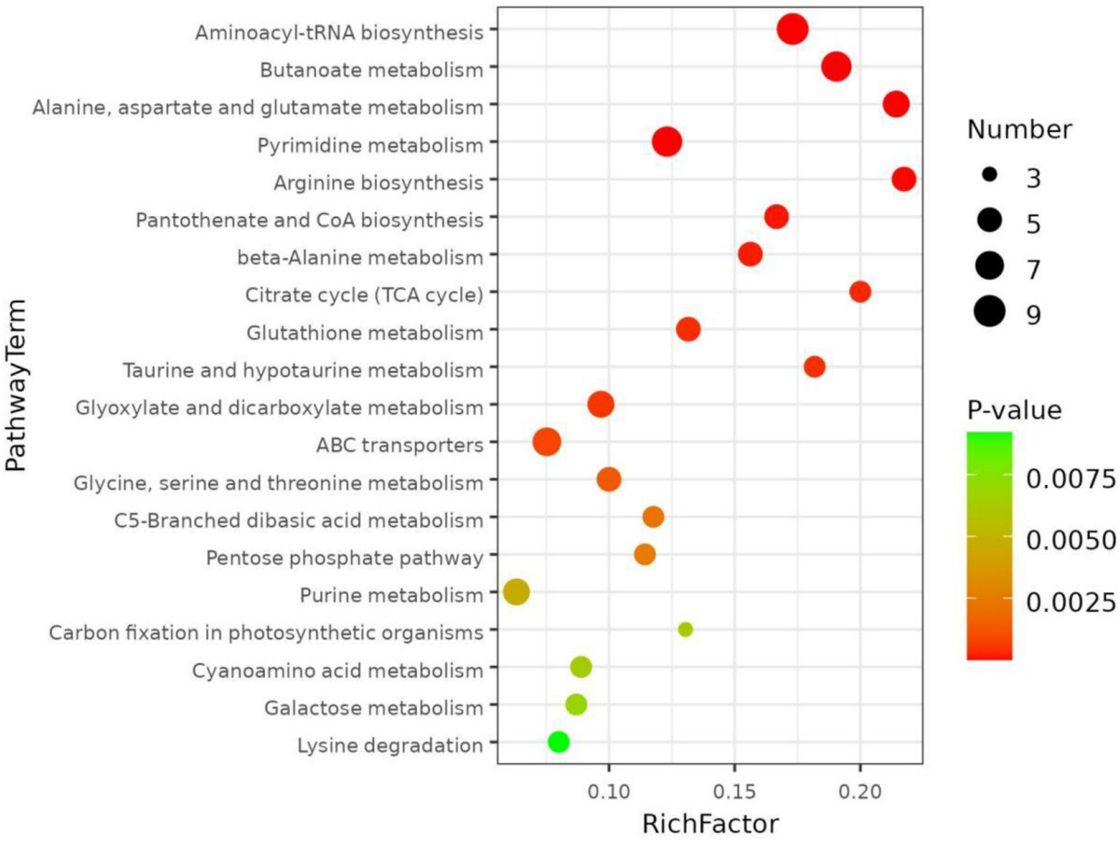

FIGURE 4
KEGG pathways enriched based on differentially abundant metabolites in different pigmented rice varieties.

criteria, precision, recall, F1 score, and accuracy, suggested that the random forest algorithm scored the highest, with an accuracy of 0.97. Logistic regression displayed the minimum accuracy of 0.89, whereas the accuracies of the XB boost (0.90) and decision tree (0.91) models were comparable (Table 1). In general, the random forest model was the optimum for predicting metabolites among different pigmented rice varieties. Several studies have also shown that a random forest can diminish overfitting and has the best separation, as it introduces randomness, possesses low noise and is applicable for complex data (16).

TABLE 1 Metric parameters of the four machine learning approaches for the discrimination of metabolites.

Model	Precision	Recall	F1-score	Accuracy
XG boost	0.902	0.900	0.898	0.90
Random forest	0.966	0.966	0.966	0.97
Logistic regression	0.898	0.888	0.888	0.89
Decision tree	0.91	0.91	0.91	0.91

3.5 Feature importance and SHAP analysis

Figure 5E shows the feature importance of the selected 7 metabolites for prediction through machine learning, which revealed that phytosphingosine and lactitol were the top 2 features. Through SHAP analysis, the influence intensities of the import eigenvalue can be compared in order. In the present study, the 7 metabolites were ranked by the average SHAP value, which denoted the scale of the influence of each variable on the simulated export. As illustrated in Figures 5F, 6, phytosphingosine and lactitol contributed to the model discriminating the different pigmented rice varieties, which was similar to the feature importance outcomes. Moreover, the seven features contributing to the prediction varied among different pigmented rice varieties, a screening method that was also employed by Zhang et al. (21).

Phytosphingosine had a greater impact on the prediction of yellow-colored rice and green-colored rice than did the other three rice varieties. Lactitol and L-tryptophan had greater impacts on the prediction of red-colored rice than the other four rice varieties did (Figure 6). D-glucose and D-mannonic acid had greater impacts on the prediction of black-colored rice than did the other four rice varieties. Maltitol had a greater impact on the prediction of purple-colored rice than on that of the other four rice varieties (Figures 5F, 6). Moreover, coupling the feature importance value with the SHPA value examination can visually show the influence

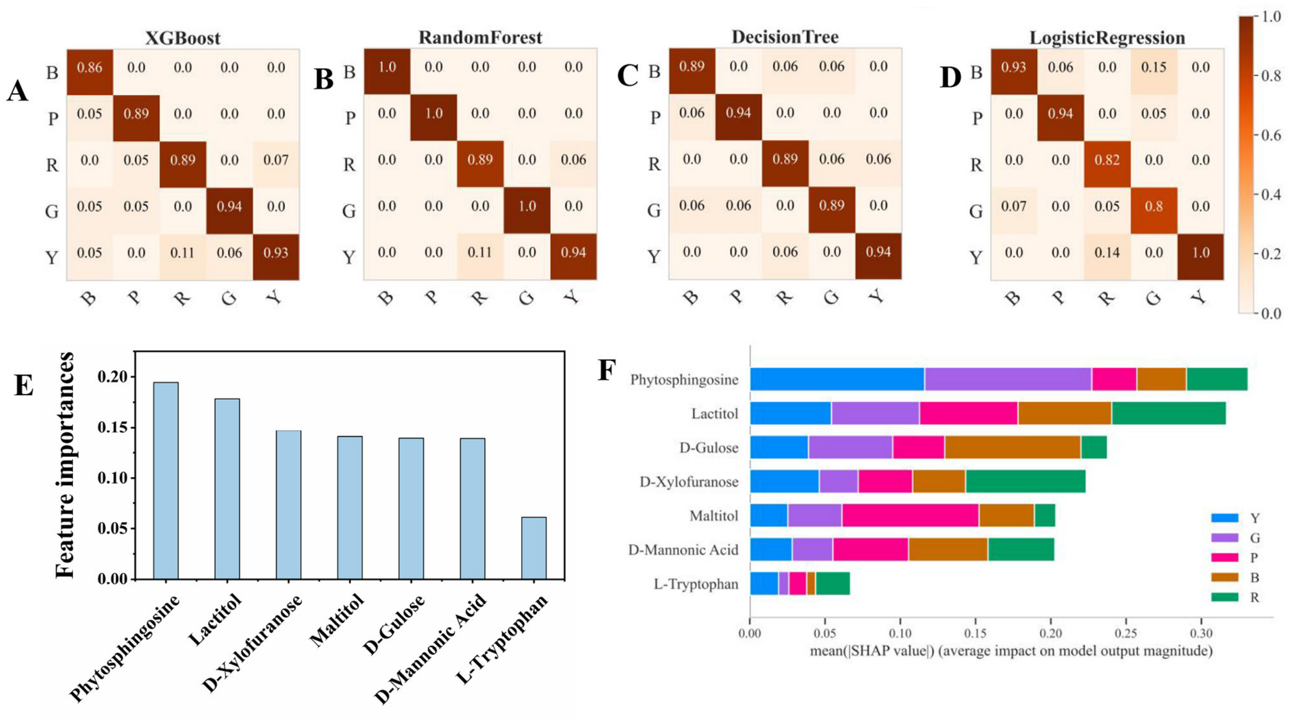

FIGURE 5
Confusion matrix and feature importance of the 7 metabolites according to the four machine learning models. **(A)** XGBoost, **(B)** random forest, **(C)** decision tree, **(D)** logistic regression, **(E)** feature importance, and **(F)** SHAP values based on random forest.

allocation for each sample, offering a probable justification for the forecast simulation and its dependence on metabolites in different pigmented rice varieties.

The long-chain base phytosphingosine is a component of sphingolipids and is present in yeast, plants and some mammalian tissues (31). Sphingosphingoid is the main component of plant biofilms and an important bioactive molecule in cells. It is involved in a variety of signal transduction pathways plays a vital role in plant growth, development and response to biotic and abiotic stresses. In the present study, it was found in pigmented rice, and it serves as an important differentially abundant metabolite.

Tryptophan (TRP) is converted into countless chemicals of biological significance, such as vitamins, and auxins (32). Fatchiyah et al. (33) used high-performance liquid chromatography (HPLC) to analyze L-tryptophan (Trp) in black rice samples and reported that Trp was the main precursor for the formation of phenolic compounds.

Soluble sugars contribute to many biological processes and structural constituents of cells (34). Kotamreddy et al. (35) found variations in glucose content in red, white and black rice via GC–MS in conjunction with metabolomics. Sugars and sugar alcohols play vital roles in the response of plants to salinity and drought stresses (36). By gas chromatography coupled with time-of-flight mass spectrometry (GC–TOF–MS), Kim et al. (37) detected three sugar alcohols in black and white rice. Figure 6 shows the boxplot profiles of the 7 features extracted from the differentially abundant metabolites in different pigmented rice varieties, which can well depict and distinguish the overall metabolomic profiles in pigmented rice. Compared with those of the other types, the levels of D-xylopyranose and D-fructose were significantly greater in black rice, indicating possible metabolic enrichment unique to black rice. D-mannonic acid was expressed at the highest level in red rice, while yellow rice also presented elevated levels, suggesting its potential involvement in oxidative stress or mannose metabolism. L-tryptophan and phosphoglycerate were predominantly elevated in purple rice, reflecting differences in amino acid and glycolytic metabolism. Notably, lactitol was almost exclusively expressed in purple rice, whereas the other types presented nearly zero levels. Similarly, maltitol levels were substantially higher in red rice and yellow rice.

The present study considered D-mannonic acid, maltitol and lactitol to be important differentially abundant metabolites, which is similar to the findings of the abovementioned reports (36, 37). Studies have shown that maltitol promotes the growth of beneficial gut bacteria, such as bifidobacteria and lactobacilli, and increases short-chain fatty acid production, enhancing gut health (38). Similarly, lactitol has been found to support the growth of these beneficial microbiota and increase the levels of SCFAs, such as butyrate and propionate, which are beneficial for gut function (39). Additionally, D-mannonic acid, a derivative of mannose metabolism, has been linked to antioxidant properties and may help modulate oxidative stress, although more direct studies are needed to fully elucidate its role in cellular protection. This finding also indicated that the key features of volatile metabolites in various colored rice varieties could be effectively extracted via GC–MS coupled with machine learning. Similar reports about features selected on the basis of metabolomics and machine learning were also published (16, 23). However, the present data and sample size were small, and the results should be expanded and validated in the future.

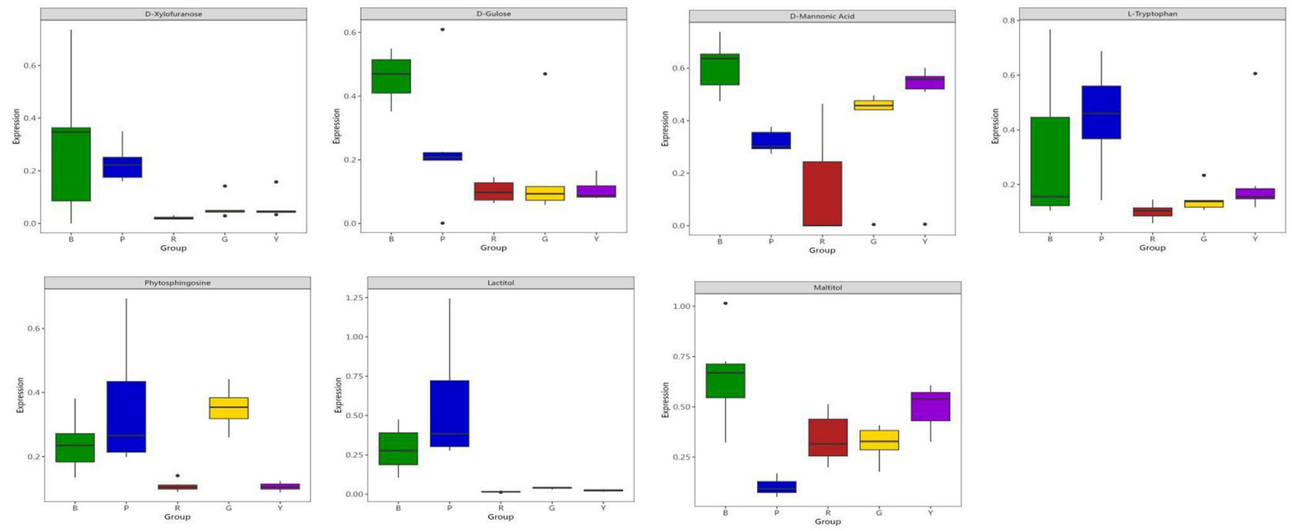

FIGURE 6
Box plot of the 7 feature metabolites in different pigmented rice varieties.

4 Conclusions

In summary, GC–MS-based metabolomics of different pigmented rice varieties was characterized, and 127 differentially abundant metabolites, which can favorably represent the majority of sample features, were screened. On the basis of metabolites with great multicollinearity above 0.8 and the chi-square test (20% feature numbers), only 7 metabolites were found to better represent the overall metabolites among the several colored rice varieties. The seven metabolites of the four machine learning models were further used for the classification of different pigmented rice varieties. The random forest model was the optimum for predicting classification, with an accuracy of 0.97. Moreover, SHAP analysis revealed that 7 metabolites can be used as potential markers for representing the metabolomic profiles. Overall, these results could provide insights into the distinctness of key gaseous metabolites in the five colored rice varieties. Machine learning approaches have proven to be rapid, useful tools for identifying key features of large data sets related to metabolomic analysis.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

KC: Methodology, Investigation, Writing – original draft, Formal analysis. RD: Software, Visualization, Data curation, Writing – original draft, Investigation. FP: Writing – original draft, Data curation, Visualization, Software, Validation. WS: Data curation, Writing – original draft, Software, Validation. LX: Software, Writing – original draft, Data curation, Validation. MZ: Validation, Writing – original draft, Data curation, Software. JG: Writing – review & editing, Supervision. RG: Supervision, Writing – review & editing. WJ: Project administration, Supervision, Funding acquisition, Writing – review & editing, Conceptualization. AA: Writing – review & editing, Supervision, Conceptualization, Funding acquisition, Project administration.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by the Foreign Expert Service Project of Shaanxi Province, China (Grant Numbers 2024WZ-YBXM-40 and 2025WZ-YBXM-45) and Scientific Plan of Shaanxi University of Technology (No. SXC-2303).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2025.1598875/full#supplementary-material

References

1. Seo WD, Kim JY, Han S-I, Ra J-E, Lee JH, Song YC, et al. Relationship of radical scavenging activities and anthocyanin contents in the 12 colored rice varieties in Korea. *J Korean Soc Appl Biol Chem.* (2011) 54:693–9. doi: 10.1007/BF03253147

2. Yodmanee S, Karrila T, Pakdeechanuan P. Physical, chemical and antioxidant properties of pigmented rice grown in Southern Thailand. *Int Food Res J.* (2011) 18:901–6. doi: 10.1080/10942912.2023.2293465

3. Callcott ET, Blanchard CL, Snell P, Santhakumar AB. The anti-inflammatory and antioxidant effects of pigmented rice consumption in an obese cohort. *Food Funct.* (2019) 10:8016–25. doi: 10.1039/C9FO02261A

4. Tiozon RJN, Sartagoda KJD, Fernie AR, Sreenivasulu N. The nutritional profile and human health benefit of pigmented rice and the impact of post-harvest processes and product development on the nutritional components: a review. *Crit Rev Food Sci Nutr.* (2023) 63:3867–94. doi: 10.1080/10408398.2021.1995697

5. Sirilertpanich P, Ekkaphan P, Andriyas T, Leksungnoen N, Ruengphayak S, Vanavichit A, et al. Metabolomics study on the main volatile components of Thai colored rice cultivars from different agricultural locations. *Food Chem.* (2024) 434:137424. doi: 10.1016/j.foodchem.2023.137424

6. Jin W, Zhang Z, Zhao S, Liu J, Gao R, Jiang P. Characterization of volatile organic compounds of different pig-mented rice after puffing based on gas chromatography-ion migration spectrometry and chemometrics. *Food Res Int.* (2023) 169:112879. doi: 10.1016/j.foodres.2023.112879

7. Qi S, He J, Han H, Zheng H, Jiang H, Hu CY, Zhang Z, Li X. Anthocyanin-rich extract from black rice (*Oryza sativa* L. Japonica) ameliorates diabetic osteoporosis in rats. *Food Funct.* (2019) 10:5350–60. doi: 10.1039/C9FO00681H

8. Chen S, Qin W, Guo Z, Li R, Ding C, Zhang S, Tan Z. Metabonomics study of fresh bruises on an apple using the gas chromatography–mass spectrometry (GC–MS) method. *Eur Food Res Technol.* (2020) 246:201–12. doi: 10.1007/s00217-019-03386-x

9. Jin W, Liu J, Zhao P, Chen X, Han H, Pei J, et al. Analysis of volatile flavor components in cooked unpolished rice of different colors from Yangxian County by headspace-gas chromatography-ion mobility spectroscopy. *Food Sci.* (2022) 43:258–64. doi: 10.7506/spkx1002-6630-20210927-324

10. Putri SP, Ikram MMM, Sato A, Dahlan HA, Rahmawati D, Ohto Y, et al. Application of gas chroma-tography-mass spectrometry-based metabolomics in food science and technology. *J Biosci Bioeng.* (2022) 133:425–35. doi: 10.1016/j.jbiosc.2022.01.011

11. Yamada T, Kamiya M, Higuchi M. Gas chromatography–mass spectrometry-based metabolomic analysis of Wagyu and Holstein beef. *Metabolites.* (2020) 10:95. doi: 10.3390/metabo10030095

12. Vieira MB, Faustino MV, Lourenço TF, Oliveira MM. DNA-based tools to certify authenticity of rice varie-ties—an overview. *Foods.* (2022) 11:258. doi: 10.3390/foods11030258

13. Fu TX, Feng YC, Zhang LY, Li X, Wang CY. Metabonomics study on rice from different geographical areas based on gas chromatography-mass spectrometry. *Food Sci.* (2019) 40:176–81. doi: 10.7506/spkx1002-6630-20180621-412

14. Zhang L, Cui D, Ma X, Han B, Han L. Comparative analysis of rice reveals insights into the mechanism of colored rice via widely targeted metabolomics. *Food Chem.* (2023) 399:133926. doi: 10.1016/j.foodchem.2022.133926

15. Wang T, An J, Chai M, Zhu Z, Jiang Y, Huang X, et al. Volatile metabolomics reveals the characteristics of the unique flavor substances in oats. *Food Chem.* (2023) 20:101000. doi: 10.1016/j.fochx.2023.101000

16. Zheng X, Pan F, Naumovski N, Wei Y, Wu L, Peng W, et al. Precise prediction of metabolites patterns using machine learning approaches in distinguishing honey and sugar diets fed to mice. *Food Chem.* (2024) 430:136915. doi: 10.1016/j.foodchem.2023.136915

17. Peng Y, Zheng C, Guo S, Gao F, Wang X, Du Z, et al. Metabolomics integrated with machine learning to discriminate the geographic origin of Rougui Wuyi rock tea. *NPJ Sci Food.* (2023) 7:7. doi: 10.1038/s41538-023-00187-1

18. Xiong Z, Feng W, Xia D, Zhang J, Wei Y, Li T, et al. Distinguishing raw pu-erh tea pro-duction regions through a combination of HS-SPME-GC-MS and machine learning algorithms. *LWT.* (2023) 185:115140. doi: 10.1016/j.lwt.2023.115140

19. Han H, Liu C, Gao W, Li Z, Qin G, Qi S, et al. Anthocyanins are converted into anthocyanidins and phenolic acids and effectively absorbed in the jejunum and ileum. *J Agric Food Chem.* (2021) 69:992–1002. doi: 10.1021/acs.jafc.0c07771

20. Runqing L. *Study on brewing characteristics, flavor characteristics, and quality of huangjiu made from different pigmented rice varieties* (Thesis for Master degree). Shaanxi University of Technology, Hanzhong, China (2024). p. 18–20.

21. Zhang X, Lu X, He C, Chen Y, Wang Y, Hu L, et al. Characterizing and decoding the dynamic alterations of volatile organic compounds and non-volatile metabolites of dark tea by solid-state fermentation with *Penicillium polonicum* based on GC–MS, GC-IMS, HPLC, E-nose and E-tongue. *Food Res Int.* (2025) 209:116279. doi: 10.1016/j.foodres.2025.116279

22. Cheng K, Xiao J, He J, Yang R, Pei J, Jin W, et al. Unraveling volatile metabolites in pigmented onion (*Allium cepa* L.) bulbs through HS-SPME/GC-MS-based metabolomics and machine learning. *Front Nutr.* (2025) 12:1582576. doi: 10.3389/fnut.2025.1582576

23. Guo T, Pan F, Cui Z, Yang Z, Chen Q, Zhao L, et al. FAPD: an astringency threshold and astringency type prediction database for flavonoid compounds based on machine learning. *J Agric Food Chem.* (2023) 71:4172–83. doi: 10.1021/acs.jafc.2c08822

24. Xiao Y, ChenH, Chen Y, Ho C-T, Wang Y, Cai T, et al. Effect of inoculation with different *Eurotium cristatum* strains on the microbial communities and volatile organic compounds of Fu brick tea. *Food Res Int.* (2024) 197:115219. doi: 10.1016/j.foodres.2024.115219

25. Ch R, Chevallier O, McCarron P, McGrath TF, Wu D, Nguyen Doan Duy L, et al. Metabolomic fingerprinting of volatile organic compounds for the geographical discrimination of rice samples from China, Vietnam and India. *Food Chem.* (2021) 334:127553. doi: 10.1016/j.foodchem.2020.127553

26. Ibba M, Söll D. Aminoacyl-tRNA synthesis. *Ann Rev Biochem.* (2000) 69:617–50. doi: 10.1146/annurev.biochem.69.1.617

27. Jumpa T, Phetcharaburanin J, Suksawat M, Pattanagul W. Physiological traits and metabolic profiles of contrasting rice cultivars under mild salinity stress during the seedling stage. *Notulae Botanicae Horti Agrobot Cluj-Napoca.* (2023) 51:13211. doi: 10.15835/nbha51213211

28. Duan R, Lin Y, Yang L, Zhang Y, Hu W, Du Y, et al. Effects of antimony stress on growth, structure, enzyme activity and metabolism of Nipponbare rice (*Oryza sativa* L.) roots. *Ecotoxicol. Environ. Safety.* (2023) 249:114409. doi: 10.1016/j.ecoenv.2022.114409

29. Liu Y, Liu J, Liu M, Liu Y, Strappe P, Sun H, et al. Comparative non-targeted metabolomic analysis reveals insights into the mechanism of rice yellowing. *Food Chem.* (2020) 308:125621. doi: 10.1016/j.foodchem.2019.125621

30. Sew YS, Aizat WM, Zainal-Abidin R-A, Ab Razak MS, Simoh S, Abu-Bakar N. Proteomic variability and nu-trient-related proteins across pigmented and non-pigmented rice grains. *Crops.* (2023) 3:63–77. doi: 10.3390/crops3010007

31. Kondo N, Ohno Y, Yamagata M, Obara T, Seki N, Kitamura T, et al. Identification of the phytosphingosine metabolic pathway leading to odd-numbered fatty acids. *Nat Commun.* (2014) 5:5338. doi: 10.1038/ncomms6338

32. Setyaningsih W, Saputro IE, Palma M, Barroso CG. Optimization of the ultrasound-assisted extraction of tryptophan and its derivatives from rice (*Oryza sativa*) grains through a response surface methodology. *J Cereal Sci.* (2017) 75:192–7. doi: 10.1016/j.jcs.2017.04.006

33. Fatchiyah F, Sari DRT, Safitri A, Cairns JR. Phytochemical compound and nutritional value in black rice from Java Island, Indonesia. *Syst Rev Pharm.* 11:414–21. doi: 10.31838/srp.2020.7.61

34. Khan N, Ali S, Zandi P, Mehmood A, Ullah S, Ikram M. Role of sugars, amino acids and organic acids in improving plant abiotic stress tolerance. *Pak J Bot.* (2020) 52:355–63. doi: 10.30848/PJB2020-2(24)

35. Kotamreddy JNR, Hansda C, Mitra A. Semi-targeted metabolomic analysis provides the basis for enhanced antiox-idant capacities in pigmented rice grains. *J Food Meas Charact.* (2020) 14:1183–91. doi: 10.1007/s11694-019-00367-2

36. Singh M, Kumar J, Singh S, Singh VP, Prasad SM. Roles of osmoprotectants in improving salinity and drought tolerance in plants: a review. *Rev. Environ. Sci. BioTechnol.* (2015) 14:407–26. doi: 10.1007/s11157-015-9372-8

37. Kim JK, Park S-Y, Lim S-H, Yeo Y, Cho HS, Ha S-H. Comparative metabolic profiling of pigmented rice (*Oryza sativa* L.) cultivars reveals primary metabolites are correlated with secondary metabolites. *J Cereal Sci.* (2013) 57:14–20. doi: 10.1016/j.jcs.2012.09.012

38. Thabuis C, Herbomez A-C, Desailly F, Ringard F, Wils D, Guérin-Deremaux L. Prebiotic-like effects of SweetPearl® Maltitol through changes in caecal and fecal parameters. *Food Nutr Sci.* (2012) 3:1375–81. doi: 10.4236/fns.2012.310180

39. Li XQ, Zhang XM, Wu X, Lan Y, Xu L, Meng XC, et al. Beneficial effects of lactitol on the composition of gut microbiota in constipated patients. *J Digest Dis.* (2020) 21:445–53. doi: 10.1111/1751-2980.12912