ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Comprehensive Analysis of Bile Medicines Based on UHPLC-QTOF-MS^E and Machine Learning

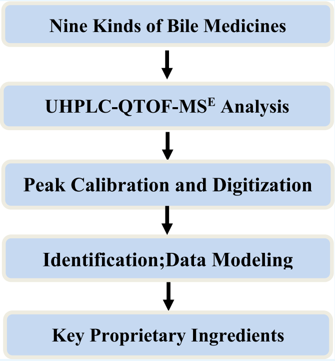
Xian rui Wang, Hao nan Wu, Ming hua Li, Xiao han Guo, Xian long Cheng,* Wen guang Jing,* and Feng Wei

Cite This: *ACS Omega* 2024, 9, 43264−43271

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Based on UHPLC-QTOF-MS^E analysis and quantized processing, combined with machine learning algorithms, data modeling was carried out to realize digital identification of bear bile powder (BBP), chicken bile powder (CIBP), duck bile powder (DBP), cow bile powder (CBP), sheep bile powder (SBP), pig bile powder (PBP), snake bile powder (SNBP), rabbit bile powder (RBP), and goose bile powder (GBP). First, 173 batches of bile samples were analyzed by UHPLC-QTOF-MS^E to obtain the retention time-exact mass (RTEM) data pair to identify bile acid-like chemical components. Then, the data were modeled by combining support vector machine (SVM), random forest (RF), artificial neural network (ANN), gradient boosting (GB), AdaBoost (AB), and Naive Bayes (NB), and the models were evaluated by the parameters of accuracy (Acc), precision (P), and area under the curve (AUC). Finally, the bile medicines were digitally identified based on the optimal model. The results showed that the RF model constructed based on the identified 12 bile acid-like chemical constituents and random forest algorithm is optimal with ACC, P, and AUC > 0.950. In addition, the accuracy of external identification verification of 42 batches of bile medicines detected at different times is 100.0%. So based on UHPLC-QTOF-MS^E analysis and combined with the RF algorithm, it can efficiently and accurately realize the digital identification of bile medicines, which can provide reference and assistance for the quality control of bile medicines. In addition, hyodeoxycholic acid, glycohyodeoxycholic acid, and taurochenodeoxycholic acid, and so forth are the most important bile acid constituents for the identification of nine bile medicines.


Nine Kinds of Bile Medicines
UHPLC-QTOF-MS^E Analysis
Peak Calibration and Digitization
Identification;Data Modeling
Key Proprietary Ingredients

■ INTRODUCTION

Animal bile medicines refer to the medicines made from the bile of animals through simple processing.[1−3] Animal bile medicines are an important part of animal medicines; their clinical use has been reported for thousands of years and has a very long history, and there are many valuable Chinese medicines, such as bear bile powder (BBP).[4−6] Compared with plant medicines, animal bile medicines have received wide-spread attention due to their strong medicinal activity, significant therapeutic effects, and low side effects. At present, the commonly used animal bile medicines in clinical practice mainly include bear bile powder (BBP), chicken bile powder (CIBP), duck bile powder (DBP), cow bile powder (CBP), sheep bile powder (SBP), pig bile powder (PBP), snake bile powder (SNBP), rabbit bile powder (RBP), and goose bile powder (GBP). The efficacy of various kinds of bile medicines is very different, and the differences between rich and poor resources are also very great. However, due to their similar characteristics and microscopic features, some production enterprises used cheap and easy-to-obtain animal bile instead of rare animal bile medicines for production cost savings, resulting in animal bile Chinese herbal medicines of varying quality, the market order is very chaotic, seriously affecting and jeopardizing the economic interests of consumers and health.[7] Therefore, it is necessary to strengthen the identification analysis and quality control of bile medicines.

To realize identification analysis and quality control of bile medicines, traditional empirical identification mainly includes macroscopic and microscopic characteristics, and physico-chemical property has been used to identity bile medicines.[7,8] Zhang et al. used chip-based nanoelectrospray ionization tandem mass spectrometry to realize rapid identification of bear bile powder from other bile sources.[9] Based on HPLC-CAD technology and chemometrics, Yuan et al. determined the content of different bile acids in BBP samples with a view to improving quality control.[10] In addition, Lei et al. constructed the recognition model for the identification of BBP and its counterfeits based on the machine learning algorithm.[11]

The abovementioned studies helped strengthen the quality control and market supervision of bile medicines to a certain

Received: September 8, 2024
Revised: September 17, 2024
Accepted: September 26, 2024
Published: October 8, 2024

https://doi.org/10.1021/acsomega.4c08260

degree. To further enrich the identification means of bile medicines, in this paper, first, we used UHPLC-QTOF-MS^E to detect nine kinds of bile medicines.[12] Further, bile acid-like chemical composition identification was performed based on chemical reference substances, and digital quantized procession was performed based on a quality control sample. Then, based on the identified chemical constituents of bile acids, the data models were constructed by combining various machine-learning algorithms to realize the digital identification and analysis of bile medicines. Finally, we selected the best model for external identification analysis based on evaluation parameters such as Acc, P, and AUC. At the same time, important variable indicator screening was performed to explore the differential key chemical components.

■ MATERIALS AND METHODS

Chemical Reference Substances and Herbal Materials. The chemical reference substances of 17 bile acids, shown in Table 2, were purchased from the National Institutes for Food and Drug Control (NIFDC) and Shanghai yuanye Bio-Technology Co., Ltd. In addition, the bile materials were collected from NIFDC and contains 37 BBPs, 35 CIBPs, 14 DBPs, 25 CBPs, 21 SBPs, 21 PBPs, 26 SNBPs, 21 RBPs, and 15 GBPs. Samples were stored in herbarium before test analysis.

Reagent Consumables and UHPLC-QTOF-MS^E Analysis. The methanol (MS grade, Lot: 10315431) was purchased from Honeywell Trading Co., Ltd. of Shanghai China. The acetonitrile (MS grade, Lot: 10315419) was purchased from Thermo Fisher Scientific shier Technology Co., Ltd. of Shanghai China. Mass spectrometry-grade ammonium formate (Lot: 102580561) was purchased from Honeywell Trading Co., Ltd. of Shanghai China. Ultrapure water (Lot: GB 19298) was purchased from Watsons Food and Beverage Co., Ltd., Guangzhou China. The 2 mL disposable sterile syringe (Lot: 20230312) was purchased from Shandong Weigao Group Medical Polymer Products Co. The Waters injection vial (Lot: 5660631710) was purchased from Waters Co., and the 0.22 μm organic filtration membrane (Lot: F210801) was obtained from Shimadzu Co.

The UHPLC-QTOF-MS^E analysis was performed on a Waters Xevo G2-XS QTof. At the same time, ESI-mode was used for detection and analysis and the MS^E data acquisition method was used, in which the data acquisition rate was set to 0.2 s; the scanning range of m/z was 100−1200; collision gas was high purity argon, and the real-time mass axis calibration solution (lock mass) was Leucine Enkephalin (LE), whose concentration is 200 ng/mL.[13] In addition, capillary: 3.0 kV; sampling cone: 40 V; source offset: 80 V; desolvation temperatures: 450 °C; desolvation gas: 600 L/h, collision energy: 10∼50 V as well as source temperatures: 120 °C.[13] Before sample analysis, the calibration of the mass axis and lock mass were performed. In addition, for the analysis of BBP, CIBP, DBP, CBP, SBP, PBP, SNBP, RBP, and GBP, the chromatographic separations were conducted on Waters Acquity UPLC BEH-C_{18} (2.1 mm × 100 mm, 1.7 μm) chromatographic column. The column temperature and injection volume were 35 °C and 2.0 μL. The gradient elution of mobile phases is shown in Table 1, in which A-methanol, B-acetonitrile, and C-5% ammonium formate solution.[13]

Sample Pretreatment. 10.00 mg of each chemical reference substance was accurately weighed, and a mixed standard solution was made at a concentration of 500 ng/mL in a suitable amount of methanol.[13] At the same time, for bile medicines, 25.00 mg of bile medicines was accurately weighed and added to 25.00 mL of methanol for ultrasonication for 30 min (power: 500 W, frequency: 40 kHz); then, the samples were cooled to room temperature and filtered through the 0.22 μm filtration membrane.[12,13] In addition, the quality control sample is a mixed sample of nine bile medicines.

Table 1. Gradient Elution Program of Liquid Chromatography[13]

time	flow (mL/min)	% A	% B	% C
0	0.3	14.0	23.0	63.0
15		24.0	29.0	47.0
20		20.0	29.0	51.0
25		15.0	29.0	56.0
30		30.0	35.0	35.0
40		25.0	72.0	3.0
41		14.0	23.0	63.0
45		14.0	23.0	63.0

Data Processing and Analysis. Based on chemical reference substances, literature research, and database comparison, we initially identified some bile acid chemical constituents. On the other hand, the mass spectrometry information of BBP, CIBP, DBP, CBP, SBP, PBP, SNBP, RBP, and GBP was processed by Progenesis QI software (Version 2.3).[14] Furthermore, the quantized data of identified bile acid chemical components were screened out, and the data models were constructed by combining SVM, RF, ANN, GB, AB and NB algorithms in Orange software (Version 3.35.0). Finally, the best model was selected for external identification analysis based on evaluation parameters such as Acc, P, and AUC. At the same time, important variable indicator screening was performed to explore the differential key chemical components.

■ RESULTS AND DISCUSSION

UHPLC-QTOF-MS^E Analysis. Under the united experimental conditions, the chromatograms of blank, mixed chemical reference substances, samples, and the quality control sample are shown in Figure 1.

As shown in Figure 1, different bile medicines showed different base-peak chromatograms as a whole. For a small number of samples, chromatogram comparison is sufficient to realize identification. However, as the number of samples increases and individual differences come to the fore, chromatogram comparison is to be inefficient. The identification efficiency can be greatly improved if an identification model can be constructed based on stable and defined chemical composition data and machine learning algorithms.[11] Therefore, in the subsequent section, we initially identified several bile acid components and quantified the base-peak chromatogram to conduct data modeling. In addition, as shown in Figure 1B and Table 2, there were 17 chemical reference substances, which can be used to identity bile acids.

Further, based on the abovementioned bile acid chemical reference substances, literature references and database comparisons,[9,10,12,15,16] we identified the bile acid components. For example, compound A (11.06_514.2830 m/z) and its dimerization (11.06_1029.5717 m/z $[2M\text{-}H]^-$) can be detected in all bile medicines. Its corresponding chemical reference substance is taurocholic acid (Rt = 11.06 min_m/z 514.2769, TCA). Using the same method, we preliminary

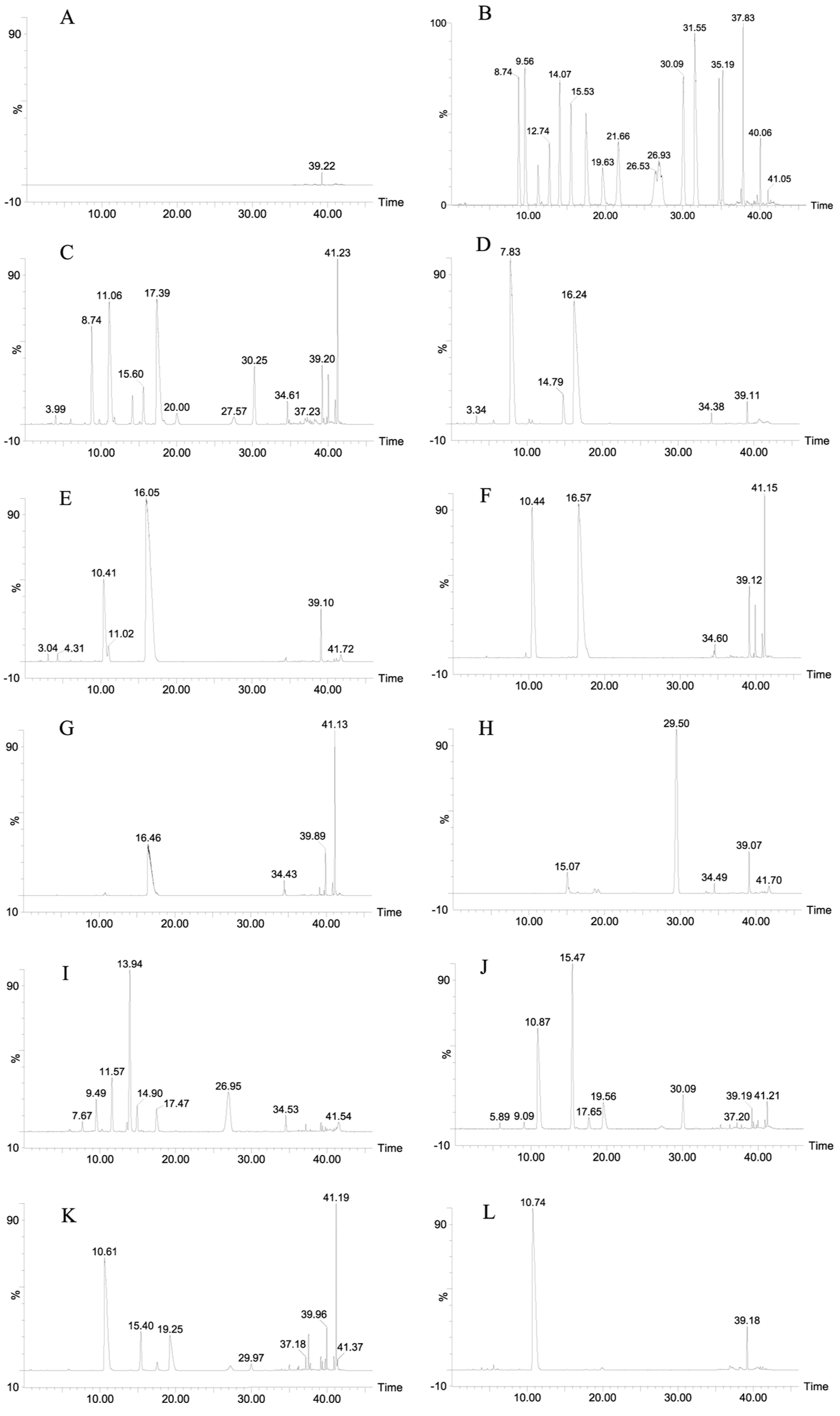


Figure 1. Base-peak chromatogram of blank, mix reference standards, some samples, and the quality control sample (A: blank; B: mixed chemical reference substances; C: QC sample; D: BBP; E: CIBP; F: DBP; G: GBP; H: RBP; I: PBP; J: CBP; K: SBP; L: SNBP).

Table 2. Specific Information on 17 Bile Acid Components

compositions	abbreviation	ions	compositions	abbreviation	ions
tauroursodeoxycholic acid	TUDCA	8.74 min_*m*/*z* 498.2922	hyodeoxycholic acid	HDCA	26.80 min_*m*/*z* 437.2958
taurohyodeoxycholic acid	THDCA	9.56 min_*m*/*z* 498.2971	glycochenodeoxycholic acid	GCDCA	26.93 min_*m*/*z* 448.3023
taurocholic acid	TCA	11.06 min_*m*/*z* 514.2769	cholic acid	CA	27.23 min_*m*/*z* 407.2870
glycoursodeoxycholic acid	GUDCA	12.74 min_*m*/*z* 498.3023	glycodeoxycholic acid	GDCA	30.09 min_*m*/*z* 448.3023
glycohyodeoxycholic acid	GHDCA	14.07 min_*m*/*z* 448.3023	taurolithocholic acid	TLCA	31.55 min_*m*/*z* 482.2889
glycocholic acid	GCA	15.53 min_*m*/*z* 464.2968	chenodeoxycholic acid	CDCA	34.70 min_*m*/*z* 437.2858
taurochenodeoxycholic acid	TCDCA	17.05 min_*m*/*z* 498.2941	deoxycholic acid	DCA	35.19 min_*m*/*z* 391.2844
taurodeoxycholic acid	TDCA	19.63 min_*m*/*z* 498.2941	lithocholic acid	LCA	37.83 min_*m*/*z* 448.2978
ursodeoxycholic acid	UDCA	21.66 min_*m*/*z* 437.2958			

Table 3. Detailed Information of 12 Bile Acid-like Chemical Constituents

composition	molecular	ions	ionic forms	sources
tauroursodeoxycholic acid	$C_{26}H_{45}NO_6S$	8.74_498.2922 *m*/*z*	$[M-H]^-$	BBP, PBP
taurocholic acid	$C_{26}H_{45}NO_7S$	11.06_514.2769 *m*/*z*	$[M-H]^-$	All Bile medicines
glycohyodeoxycholic acid	$C_{26}H_{43}NO_5$	14.07_448.3023 *m*/*z*	$[M-H]^-$	PBP
glycocholic acid	$C_{26}H_{43}NO_6$	15.53_464.2968 *m*/*z*	$[M-H]^-$	CBP, SBP, RBP
taurochenodeoxycholic acid	$C_{26}H_{45}NO_6S$	17.05_498.2941 *m*/*z*	$[M-H]^-$	All Bile medicines
taurodeoxycholic acid	$C_{26}H_{44}NO_6S$	19.63_498.2941 *m*/*z*	$[M-H]^-$	CBP, RBP, SBP
hyodeoxycholic acid	$C_{24}H_{40}O_4$	26.80_437.2958 *m*/*z*	$[M + HCOO]^-$	PBP
glycochenodeoxycholic acid	$C_{26}H_{43}NO_5$	26.93_448.3023 *m*/*z*	$[M-H]^-$	CBP, PBP
glycodeoxycholic acid	$C_{26}H_{43}NO_5$	30.09_448.3023 *m*/*z*	$[M-H]^-$	CBP, RBP, SBP
taurolithocholic acid	$C_{26}H_{45}NO_5S$	31.55_482.2889 *m*/*z*	$[M-H]^-$	CBP, PBP
chenodeoxycholic acid	$C_{24}H_{40}O_4$	34.70_437.2858 *m*/*z*	$[M + HCOO]^-$	BBP, CIBP, PBP, DBP, GBP
deoxycholic acid	$C_{24}H_{40}O_4$	35.19_391.2844 *m*/*z*	$[M-H]^-$	CBP, SBP

identified 12 bile acid-like chemical constituents in bile medicines. The detailed information on 12 bile acid-like chemical constituents is shown in Table 3.

At the same time, in UHPLC-QTOF-MS^E analysis of BBP, CIBP, DBP, CBP, SBP, PBP, SNBP, RBP, and GBP, we explored sample pretreatment and acquisition methods. The results showed that ultrasonic extraction (power: 500 W, frequency: 40 kHz) with methanol for 30 min has the better extraction effect. In addition, the MS^E mode was used to obtain the mass spectrometry information, thus ensuring more data information.[13,17] However, the reality of bile medicines is that they were largely difficult to fragment and were in the form of parent ions, which is consistent with what is documented in the database.[13,15]

■ DATA MODELING AND DISCUSSION

The Progenesis QI software (version 2.3) accomplished the quantized data transformation of bile medicines using mixed QC samples as standards. Further, the quantized data of 12 bile acid components that had been identified in 173 batches of samples were screened out and finally imported into Orange software (Version 3.35.0) and combined with machine learning algorithms to construct data models. The best model was filtered out for external validation and identification of 42 batches of samples. At the same time, the detailed information on 173 batches of sample data used for data modeling is shown in Table S1.

The 12 bile acid chemical components screened were used as data variables. The data model was constructed by combining SVM, RF, NB, GB, AB, and ANN, and 10-fold cross-validation was performed, and the evaluation parameters such as accuracy (Acc), F1-scores, AUC, Recall, and precision (P) are shown in Table 4.

Table 4. Evaluation Parameters for Different Models in 10-Fold Cross-Validation

models	AUC	Acc	F1-score	*P*	recall
RF	0.999	0.960	0.959	0.961	0.960
ANN	0.997	0.896	0.877	0.871	0.896
NB	0.998	0.879	0.882	0.923	0.879
SVM	0.993	0.850	0.814	0.803	0.850
GB	0.985	0.838	0.793	0.767	0.838
AB	0.970	0.948	0.948	0.953	0.948

In the model evaluation of machine learning, Acc, AUC, and P are three important indicators. Acc represents the accuracy of model identification, and AUC specifically refers to the area under the ROC curve. The larger the value (0.5−1.0), the better the effect of the classifier, and the P represents the proportion of positive samples predicted by the model to be positive samples that are actually positive samples, which also reflects the classification effect of the model, especially if the data are unevenly distributed.[18−20] As shown in Table 4, as far as the evaluation parameter-AUC is concerned, the AUC of each model was greater than 0.970 (max: 1.000) and was on the same order of magnitude. It suggests that the overall classification results for each model are good. Further, in terms of Acc, the Acc of all the models is not less than 0.835; among them, the GB model has the smallest Acc of 0.838, while the RF model has the largest accuracy of 0.960. At the same time, the Accs of ANN, AB, and NB models are 0.896, 0.948, and 0.879. Therefore, the RF model has the highest discrimination accuracy. Since the sample size of bile medicines was not homogeneous, the precision is equally informative. In terms of P, RF > 0.960 > AB > 0.950 > NB > 0.900 > ANN > 0.850 > SVM > 0.800 > GB. On the other hand, recall refers to the proportion of all samples that are actually positive samples predicted by the model as positive samples. Therefore, the recall is more concerned with the degree of coverage of positive samples. The F1-score is an indicator that comprehensively considers precision and recall and is a harmonic average. The higher the F1-score, the better the performance of the model. As shown in Table 4, the RF model has a recall and F1-score of 0.960 and 0.959, higher than the recall and F1-score of other models. In summary, compared to SVM, NB, GB, AB, and ANN models, the RF model has the best identification effect and is the optimal model. It can effectively realize the digital identification and analysis of bile medicines based on the RF identification model that was constructed based on the quantized data of 12 bile acid components and the RF algorithm.

Meanwhile, we adopt an automatic optimization strategy in data modeling. The specific parameters of each model are shown in Table 5.

Table 5. Specific Parameters of Each Model

models	specific parameters
RF	number of tree ≤10; maximal number of considered features: unlimited; maximal tree depth: unlimited; stop splitting nodes with maximum instance ≤3
ANN	neurons in hidden layers: 100; activation: ReLu; solver: Adam; number of iterations ≤200
NB	delete empty column: yes; discretizes numeric values: 4; equal frequency: yes
SVM	cost (C): 1.00; regression loss epsilon (ε): 0.30; kernel: RBF; iteration limit ≤100
AB	base estimator: tree; number of estimators ≤50; learning rate: 1.00; classification algorithm: SAMME.R; regression loss function: linear
GB	method: catboost; number of trees ≤10; learning rate: 0.10; limit depth of individual trees ≤3; fraction of features for each tree: 1.00

On the other hand, under the same conditions, we also compared data models constructed based on all quantized data and the models constructed based on quantized data for 12 bile acid-like chemical components. The evaluation parameters of each model constructed based on all quantized data are shown in Table 6.

Table 6. Evaluation Parameters for Different Models Based on all Quantized Data

models	AUC	Acc	F1-score	*P*	recall
RF	0.993	0.928	0.924	0.931	0.928
ANN	0.852	0.836	0.827	0.854	0.836
NB	0.983	0.695	0.720	0.883	0.695
SVM	0.901	0.698	0.707	0.731	0.698
GB	0.980	0.818	0.775	0.840	0.818
AB	0.965	0.939	0.939	0.942	0.939

Comparing Table 4 and Table 6, it can be found that the evaluation parameters of each model constructed using the quantized data of 12 identified bile acid-like chemical constituents have been significantly improved compared to each model constructed using all of the quantized data; for example, the RF model has an increase of 0.032 in the Acc and recall. The SVM model has an increase of 0.152 in the Acc and recall. Therefore, not all quantized data are valid and reliable data in data modeling of bile medicines. The identified quantized data of 12 cholic acid components are combined with machine learning algorithms to build data models, which can reduce interference of irrelevant data, thus improving the model identification effect. On the other hand, all 12 bile acids have been identified and reported to be stably detected in one or more species of BBP, CIBP, DBP, CBP, SBP, PBP, SNBP, RBP, and GBP, and some of them are even proprietary chemical components. Therefore, these chemical compositions are stable and reliable and can be detected at different times and on different instruments, which facilitates the construction of data models. In addition, the quality control sample, a mixed sample of BBP, CIBP, DBP, CBP, SBP, PBP, SNBP, RBP, and GBP, was taken as the reference to perform peak position correction and data conversion. As we all know, it is difficult to directly place the mass spectrometry data of nine kinds of bile medicines in a unified analytical system, which is necessary for data modeling due to their different data volumes and characterizations. However, the quality control sample provided us with the possibility. Using a quality control sample as the reference for peak correction and data transformation, the mass spectrometry data of nine bile medicines can be integrated into a unified analysis system by converting three-dimensional (3D) LC/MS mass spectra into two-dimensional (2D) data matrices.

External Identification and Validation. According to the constructed RF model, the quantized data of the 42 batches of samples collected in different periods and not used as training sets were used as inputs for external validation identification. The detailed information on 42 batches of sample data used for appraisal verification is shown in Table S2, and the identification results and sample information are shown in Table 7.

Table 7. Digital Identification Results of Bile Medicines Based on the RF Model

bile medicines	identification results	bile medicines	identification results
DBP01	DBP	SBP03	SBP
DBP02	DBP	SBP04	SBP
DBP03	DBP	PBP01	PBP
DBP04	DBP	PBP02	PBP
BBP01	BBP	PBP03	PBP
BBP02	BBP	PBP04	PBP
BBP03	BBP	SNBP01	SNBP
BBP04	BBP	SNBP02	SNBP
BBP05	BBP	SNBP03	SNBP
BBP06	BBP	SNBP04	SNBP
BBP07	BBP	RBP01	RBP
CIBP01	CIBP	RBP02	RBP
CIBP02	CIBP	RBP03	RBP
CIBP03	CIBP	RBP04	RBP
CIBP04	CIBP	RBP05	RBP
CBP01	CBP	RBP06	RBP
CBP02	CBP	GBP01	GBP
CBP03	CBP	GBP02	GBP
CBP04	CBP	GBP03	GBP
SBP01	SBP	GBP04	GBP
SBP02	SBP	GBP05	GBP

As shown in Table 7, after identification verification, 42 batches of bile medicines (4 batches of DBP; 7 batches of BBP, 4 batches of CIBP, 4 batches of CBP, 4 batches of SBP, 4 batches of PBP, 4 batches of SNBP, 6 batches of RBP, and 5 batches of GBP), detected at different times, can be correctly identified and recognized in the constructed RF model, with a correct rate of 100.0%, which is consistent with the actual situation. This indicated that the RF model based on the quantized data of bile acid components and the RF algorithm has certain practical value and can effectively realize the digital identification of bile medicines.

DISCUSSION ON FITTING AND ROBUSTNESS OF RANDOM FOREST

Overfitting is a situation where a machine learning model performs well on training data but poorly on the test data. To avoid overfitting, the data from the 173 batches of bile medicines used for model construction and the 42 batches of bile medicines used for external identification and validation were detected and analyzed by different researchers in different periods, fully considering the influence of changes in external analysis conditions and personnel operation. We reduced the number of random forest trees to reduce the complexity of the model in constructing the RF model. Moreover, 10-fold cross-validation is a good way to prevent model overfitting. Therefore, in the model evaluation, we used 10-fold cross-validation, which helps to reduce overfitting as each sample is validated, resulting in a more representative evaluation of the model's performance. In addition, the external verification correct rate of the RF model for samples in different periods is as high as 100.0%, which also shows that the RF model is robust. On the other hand, 12 bile acid components, which are stable and reproducibly detectable, are taken as data variables and the quantized processing based on quality control samples also ensured the robustness of the RF model.

PRELIMINARY EXPLORATION OF KEY DIFFERENTIAL BILE ACID COMPONENTS

On the basis of the random forest model, we conducted an exploration of key differential bile acid components through the Gini index.[21,22] The Gini index is a measure of the purity of the samples in a data set and can also be used to assess the importance of variable features. In random forests, the Gini index is used to select the best features when decision tree nodes. By comparing the change in the Gini index when splitting using different features, we can assess the contribution of each variable feature to the improvement of the model purity. The smaller the Gini index, the greater the contribution of the variable feature to the model purity improvement and therefore the higher its importance. The Gini indices of the 12 bile acid components are shown in Table 8.

Table 8. Gini Indices of the 12 Bile Acid Components

composition	Gini index
glycochenodeoxycholic acid	0.244
glycocholic acid	0.227
taurocholic acid	0.220
glycodeoxycholic acid	0.216
tauroursodeoxycholic acid	0.207
taurodeoxycholic acid	0.186
chenodeoxycholic acid	0.176
taurochenodeoxycholic acid	0.167
taurolithocholic acid	0.166
deoxycholic acid	0.164
glycohyodeoxycholic acid	0.115
hyodeoxycholic acid	0.101

Generally speaking, the smaller Gini index means that the selected samples in the set are less likely to be wrongly classified, which fits with the search for proprietary chemical components in chemical analysis. As shown in Table 7, hyodeoxycholic acid has the lowest Gini index value of 0.101, and the Gini index values of glycohyodeoxycholic acid, deoxycholic acid, taurolithocholic acid, and taurochenodeoxycholic acid are 0.115, 0.164, 0.166, and 0.167, which means that the top 5 chemical components are the most important among the 12 bile acid constituents for the proprietary identification of bile acid medicines. Combined with Table 2, it can be determined that hyodeoxycholic acid and glycohyodeoxycholic acid are the proprietary chemical components of PBP samples. Deoxycholic acid can be only detected in CBP and SBP samples. Taurolithocholic acid can be only detected in CBP and PBP samples. In addition, taurochenodeoxycholic acid can be detected in all bile medicines, but its intensity (content) varies greatly among different bile medicines, so it is equally important for identifying bile medicines. Further, we found that the accuracy (0.953) and precision (0.956) of the RF model were slightly reduced when the data of the above five bile acid components were removed from the original data. In terms of modeling effects, it is further shown that these chemical marker components are helpful for the identification of bile medicines. On the other hand, It also proves that screening the proprietary chemical constituents of bile herbs is feasible based on the Gini index (small value). Unfortunately, there is no significant difference in the effect of the RF models after only removing a certain chemical composition data. The abovementioned results prompt us to use the above five bile acid components as a "composition combination" to identify bile medicines.

Research Limitations and Prospects. In this study, we realized the identification and analysis of nine kinds of bile herbs based on machine learning and UHPLC-QTOF-MS^E and preliminarily explored the differential bile acid constituents. However, it is undeniable that there are still some shortcomings. This study relies on 173 batches of sample data for modeling, and 42 batches are used for verification and testing. Although it may be enough for the initial research, a larger amount of data is necessary for further research based on machine learning, which can further enhance the reliability and generalizability of the model. As is well-known, the random forest has low interpretability, so it needs to be analyzed by a more interpretable model in the future.

CONCLUSIONS

In this paper, 9 kinds of bile medicines were analyzed by UHPLC-QTOF-MS^E, and 12 bile acid-like chemical components such as taurochenodeoxycholic acid, hyodeoxycholic acid, and glycodeoxycholic acid were successfully identified. Further, the quantized data of 12 bile acid components were combined with SVM, RF, ANN, NB, GB, and AB machine learning algorithms to construct data models, and the results showed that the random forest model was the optimal model with Acc = 0.960, P = 0.961. After external validation and identification, the random forest model achieved efficient and accurate identification of nine bile medicines with a correct rate of 100.0%. In addition, the exploration of key difference components based on the Gini index showed that hyodeoxycholic acid glycohyodeoxycholic acid, deoxycholic acid, taurolithocholic acid, and taurochenodeoxycholic acid are the key components for identifying bile medicines. Based on

UHPLC-QTOF-MS^E analysis and combined with the RF algorithm, it can efficiently and accurately realize the digital identification of bile medicines, which can provide reference and assistance for the quality control and digital identity of bile medicines.

■ ASSOCIATED CONTENT

Data Availability Statement

Mass spectra of all bile medicines were transformed by Progenesis QI software (Version 2.3); The 12 bile acid-like components were identified based on chemical control substances and references and the HMDB database; the quantized data screening of 12 bile acid-like components was done through WPS office software; data modeling, external validation analysis, and exploration of key variance components are accomplished with Orange analytics software (Version 3.35.0). The data information can be obtained in the Supporting Information.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c08260.

Detailed information on 173 batches of sample data used for data modeling and detailed information on 42 batches of sample data used for appraisal verification (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Xian long Cheng − *Institute for Control of Traditional Chinese Medicine and Ethnic Medicine, National Institutes for Food and Drug Control, Beijing 102629, China*; Email: cxl@nifdc.org.cn

Wen guang Jing − *Institute for Control of Traditional Chinese Medicine and Ethnic Medicine, National Institutes for Food and Drug Control, Beijing 102629, China*; Email: jingwenguang@nifdc.org.cn

Authors

Xian rui Wang − *Institute for Control of Traditional Chinese Medicine and Ethnic Medicine, National Institutes for Food and Drug Control, Beijing 102629, China*; orcid.org/0000-0002-9793-9596

Hao nan Wu − *Institute for Control of Traditional Chinese Medicine and Ethnic Medicine, National Institutes for Food and Drug Control, Beijing 102629, China; Faculty of Functional Food and Wine, Shenyang Pharmaceutical University, Shenyang 110016, China*

Ming hua Li − *Institute for Control of Traditional Chinese Medicine and Ethnic Medicine, National Institutes for Food and Drug Control, Beijing 102629, China*

Xiao han Guo − *Institute for Control of Traditional Chinese Medicine and Ethnic Medicine, National Institutes for Food and Drug Control, Beijing 102629, China*

Feng Wei − *Institute for Control of Traditional Chinese Medicine and Ethnic Medicine, National Institutes for Food and Drug Control, Beijing 102629, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c08260

Author Contributions

H.n.W. made the same contribution and should be the common first author. Data curation and conceptualization: X.W., H.W., and M.L.; formal analysis: W.J., M.L., and X.G.; funding acquisition: X.C. and F.W.; investigation: X.w., W.J., M.L., and X.G.; project administration: X.C. and F.W.; writing—original draft: X.W. and H.W.; writing—review and editing: X.W., H.W., W.J., X.G., and X.C.; software: H.W., X.G., W.J., M.L., X.W., X.C., and F.W.; supervision: W.J., X.C., and F.W.; and project design: X.W., W.J., X.C., and F.W.

Funding

The research was funded by Training Fund for academic leaders of National Institutes for Food and Drug Control (2023 × 10) and the National Key R&D Program of China (2023YFC3504105).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Firstly, we thank the Training Fund for academic leaders of National Institutes for Food and Drug Control (2023 × 10) and the National Key R&D Program of China (2023YFC3504105). Then we thank the Institute for Control of Traditional Chinese Medicine and Ethnic Medicine, the National Institutes for Food and Drug Control for their support, Finally, we thank the National Medical Products Administration State Key Laboratory of Drug Regulatory Science for support.

■ ABBREVIATIONS

Acc	accuracy
ANN	artificial neural network
AUC	area under the curve
BBP	bear bile powder
CA	cholic acid
CDCA	chenodeoxycholic acid
CBP	cow bile powder
CIBP	chicken bile powder
DBP	duck bile powder
DCA	deoxycholic acid
EMRT	exact mass-retention time
GBP	goose bile powder
GCA	glycocholic acid
GCDCA	glycochenodeoxycholic acid
GDCA	glycodeoxycholic acid
GHDCA	glycohyodeoxycholic acid
GUDCA	glycoursodeoxycholic acid
HDCA	hyodeoxycholic acid
LCA	lithocholic acid
NB	Naive Bayes
P	precision
PBP	pig bile powder
RBP	rabbit bile powder
RF	Random Forest
SBP	sheep bile powder
SNBP	snake bile powder
SVM	Support Vector Machine
TCA	taurocholic acid
TCDCA	taurochenodeoxycholic acid
TDCA	taurodeoxycholic acid
THDCA	taurohyodeoxycholic acid
TUDCA	tauroursodeoxycholic acid
TLCA	taurolithocholic acid
UDCA	ursodeoxycholic acid

■ REFERENCES

(1) Li, X. Y.; Su, F. F.; Jiang, C.; Zhang, W.; Wang, F.; Zhu, Q.; Yang, G. Development history and prospect of Fel Ursi. *Zhongguo Zhong Yao Za Zhi* **2022**, *47* (16), 4284−4291.

(2) Xiong, J.; Zheng, T. J.; Shi, Y.; Wei, F.; Ma, S. C.; He, L.; Wang, S. C.; Liu, X. S. Analysis of the fingerprint profile of bioactive constituents of traditional Chinese medicinal materials derived from animal bile using the HPLC-ELSD and chemometric methods: An application of a reference scaleplate. *J. Pharm. Biomed. Anal.* **2019**, *174* (174), 50−56.

(3) Tian, N.; Yuan, Y.; Jin, Y.; Yang, Q.; Zhang, T.; Li, J. D.; Wang, L.; Jiang, C.; Huang, L. Q. DNA fingerprinting identification of bile power (bile) medicines. *Zhongguo Zhong Yao Za Zhi* **2020**, *45* (5), 1064−1069.

(4) Yamaguchi, S.; Qian, Z. Z.; Nohara, T. Bile acids of Fel Ursi. *Chem. Pharm. Bull. (Tokyo)* **1998**, *46* (10), 1653−1655.

(5) Jung, H. W.; Hwang, J. H. Anticancer Effects of Ursi Fel Extract and Its Active Compound, Ursodeoxycholic Acid, in FRO Anaplastic Thyroid Cancer Cells. *Molecules* **2021**, *26* (17), 5309.

(6) Li, Y. W.; Zhu, X. Y.; But, P. P.; Yeung, H. W. Ethnopharmacology of bear gall bladder: I. *J. Ethnopharmacol.* **1995**, *47* (1), 27−31.

(7) Li, W. L.; Xing, L. H.; Xue, D. S.; Qu, H. B. An authentication method of bear bile powder based on the near infrared spectroscopy. *Guang Pu Xue Yu Guang Pu Fen Xi* **2011**, *31* (3), 673−676.

(8) Deng, M. Z.; Zhao, C. Z.; Peng, X. F.; Yang, J. C. Study on qualitation and quantitation method of Xiongdan pills. *Chin. J. Pharm. Anal.* **2012**, *32* (01), 127−131.

(9) Zhang, Y.; Wei, J.; Li, L.; Liu, Y.; Sun, S.; Xu, L.; Liu, S.; Wang, Z.; Yang, L. Rapid identification of bear bile powder from other bile sources using chip-based nano-electrospray ionization tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2022**, *36* (15), No. e9326.

(10) Yuan, M.; Zhou, T.; Lei, K.; Liu, Y.; Li, M.; Zeng, D.; Guo, Y.; Guo, L. Identification of the Authenticity and Geographical Origin of Bear Bile Powder by Using High Performance Liquid Chromatography - Charged Aerosol Detector Fingerprints Combined with Chemometrics. *Chem. Biodivers.* **2023**, *20* (3), No. e202201109.

(11) Lei, K.; Yuan, M.; Li, S.; Zhou, Q.; Li, M.; Zeng, D.; Guo, Y.; Guo, L. Performance evaluation of E-nose and E-tongue combined with machine learning for qualitative and quantitative assessment of bear bile powder. *Anal. Bioanal. Chem.* **2023**, *415* (17), 3503−3513.

(12) Shi, Y.; Xiong, J.; Sun, D.; Liu, W.; Wei, F.; Ma, S.; Lin, R. Simultaneous quantification of the major bile acids in artificial Calculus bovis by high-performance liquid chromatography with precolumn derivatization and its application in quality control. *J. Sep. Sci.* **2015**, *38* (16), 2753−2762.

(13) Wang, X.; Wu, H.; Li, M.; Guo, X.; Cheng, X.; Jing, W.; Wei, F. A Comprehensive Analysis of Fel Ursi and Its Common Adulterants Based on UHPLC-QTOF-MS^E and Chemometrics. *Molecules* **2024**, *29* (13), 3144.

(14) Zhang, J.; Yang, W.; Li, S.; Yao, S.; Qi, P.; Yang, Z.; Feng, Z.; Hou, J.; Cai, L.; Yang, M.; Wu, W.; Guo, D. A. An intelligentized strategy for endogenous small molecules characterization and quality evaluation of earthworm from two geographic origins by ultra-high performance HILIC/QTOF MS(E) and Progenesis QI. *Anal. Bioanal. Chem.* **2016**, *408* (14), 3881−3890.

(15) Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B.; et al. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* **2022**, *50* (D1), D622−D631.

(16) Chen, D.; Lin, S.; Xu, W.; Huang, M.; Chu, J.; Xiao, F.; Lin, J.; Peng, J. Qualitative and Quantitative Analysis of the Major Constituents in Shexiang Tongxin Dropping Pill by HPLC-Q-TOF-MS/MS and UPLC-QqQ-MS/MS. *Molecules* **2015**, *20* (10), 18597−18619.

(17) Zheng, W.; Gao, R.; Wang, F.; Shan, G.; Gao, H. Identification of Chemical Constituents in Zhizhu Pills Based on UPLC-QTOF-MS^E. *J. AOAC Int.* **2022**, *105* (6), 1555−1575.

(18) Deo, R. C. Machine Learning in Medicine. *Circulation* **2015**, *132* (20), 1920−1930.

(19) Choi, R. Y.; Coyner, A. S.; Kalpathy-Cramer, J.; Chiang, M. F.; Campbell, J. P. Introduction to Machine Learning, Neural Networks, and Deep Learning. *Transl. Vis. Sci. Technol.* **2020**, *9* (2), 14.

(20) Gong, Y.; Ding, W.; Wang, P.; Wu, Q.; Yao, X.; Yang, Q. Evaluating Machine Learning Methods of Analyzing Multiclass Metabolomics. *J. Chem. Inf. Model.* **2023**, *63* (24), 7628−7641.

(21) Biró, T. S.; Néda, Z. Gintropy: Gini Index Based Generalization of Entropy. *Entropy (Basel)* **2020**, *22* (8), 879.

(22) Zhang, Y.; Nie, B.; Du, J.; Chen, J.; Du, Y.; Jin, H.; Zheng, X.; Chen, X.; Miao, Z. Feature selection based on neighborhood rough sets and Gini index. *PeerJ Comput. Sci.* **2023**, *9*, No. e1711.