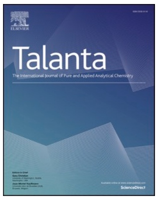
Contents lists available at ScienceDirect

Talanta

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



Using random forest to detect multiple inherited metabolic diseases simultaneously based on GC-MS urinary metabolomics

Nan Chen [a], Hai-Bo Wang [a], Ben-Qing Wu [b], Jian-Hui Jiang [a,*], Jiang-Tao Yang [c], Li-Juan Tang [a,**], Hong-Qin He [d], Dan-Dan Linghu [d]

[a] State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha, 410082, China
[b] Department of Pediatric, University of Chinese Academy of Sciences-Shenzhen Hospital, Shenzhen, 518000, PR China
[c] Shenzhen Aone Medical Laboratory Co, Ltd, Shenzhen, 518000, PR China
[d] Yuncheng Maternal and Child Health Hospital, Yuncheng, Shanxi, 044000, PR China

ARTICLE INFO

Keywords:
Inherited metabolic diseases
GC-MS
Random forest
Metabolomics analysis
Multi-classification

ABSTRACT

Inborn errors of metabolism, also known as inherited metabolic diseases (IMDs), are related to genetic mutations and cause corresponding biochemical metabolic disorder of newborns and even sudden infant death. Timely detection and diagnosis of IMDs are of great significance for improving survival of newborns. Here we propose a strategy for simultaneously detecting six types of IMDs via combining GC-MS technique with the random forest algorithm (RF). Clinical urine samples from IMD and healthy patients are analyzed using GC-MS for acquiring metabolomics data. Then, the RF model is established as a multi-classification tool for the GC-MS data. Compared with the models built by artificial neural network and support vector machine, the results demonstrated the RF model has superior performance of high specificity, sensitivity, precision, accuracy, and matthews correlation coefficients on identifying all six types of IMDs and normal samples. The proposed strategy can afford a useful method for reliable and effective identification of multiple IMDs in clinical diagnosis.

1. Introduction

Inborn errors of metabolism (IEMs), known as inherited metabolic diseases (IMDs), are a special group of rare diseases in newborns. Such diseases can threaten multiple systems of children's body, including the nervous system, digestive system, circulatory system, metabolic system and so on. Most of the pathophysiological changes of IMDs can affect human organs directly or indirectly, especially the development and function of the brain [1,2]. They may lead to disability and even sudden infant death. IMDs are very urgent and the clinical manifestations are mostly nonspecific at the beginning [3]. The body has been irreversibly damaged when symptoms appear. For IMDs, the earlier we start the treatment, the smaller the damage to the body. Timely detection and diagnosis of IMDs are of great significance for newborns [4–6].

In the 1960s, the concept of newborn screening was introduced when Dr. Guthrie conducted bacterial inhibition assay for phenylketonuria [7–9]. Over time, newborn screening has been developed from a bacterial inhibition assay to some complex methods. Nowadays, it refers to the screening of some congenital and genetic diseases, which endanger children's lives and cause children's physical and intellectual development obstacles, to make early diagnosis using fast, simple and sensitive testing methods. Laboratory diagnosis methods for IMDs include enzyme analysis, metabolite determination, gene microarray and high-throughput sequencing [10,11]. Due to the high cost of genetic testing and the high-quality requirements of specimens for enzyme analysis, the most commonly used screening methods in clinical practice are developed on gas chromatography-mass spectrometry (GC-MS) or chromatography-tandem mass spectrometry (GC-MS/MS) [12]. Among them, GC-MS is the most sensitive technique for the diagnosis of organic aciduria. With the development of urease pretreatment technology, GC-MS detection is not limited to organic aciduria, but also widely applied for diagnosing abnormal amino acid metabolism. The GC-MS technology has been to as an important method for screening complex organic aciduria and abnormal amino acid metabolism [13–16].

The metabolomics data generated using GC-MS is generally too complex for the treatment of traditional statistical methods [17]. There

* Corresponding author.
** Corresponding author.
E-mail addresses: jianhuijiang@hnu.edu.cn (J.-H. Jiang), tanglijuan@hnu.edu.cn (L.-J. Tang).

https://doi.org/10.1016/j.talanta.2021.122720
Received 30 April 2021; Received in revised form 15 July 2021; Accepted 16 July 2021
Available online 19 July 2021

are a large number of endogenous small molecules with physiological effects and functions in organisms. Biomarkers and functional substances with specific research significance are only a few specific objects. In the context of the entire incident, a small number of functional objects will be relatively severely interfered by useless objects, resulting in high noise in the metabolomics data. Generally speaking, the number of metabolites detected by non-targeted metabolome is much greater than the number of samples. So, traditional statistical methods cannot be used to process metabolomics data having such a feature of high dimensionality. There are also various factors making it difficult to identify and qualitatively analyze metabolomics data, such as isomers, metabolites with similar physical and chemical properties, liquid phase system and so on. Moreover, the distribution of metabolomics data is very irregular, and there may be many zero values in the data, which requires more complex and reasonable statistical analysis strategies to reveal the hidden complex data relationships. Therefore, it becomes critical in metabolomics research upon how to extract valuable information from metabolomics data and construct reasonable data models [18,19].

Here we propose combining GC-MS with an ensemble learning algorithm of random forest (RF) for identifying multiple types of IMDs from complex metabolomics data. Ensemble learning is a strategy that integrates multiple models to circumvent the inherent defects and limitations of single model by letting the models learn from each other and enhance the performance. As a typical ensemble learning algorithm, RF is a combination of a set of tree predictors, of which each tree grows depending on randomly selected samples and random combinations of vectors, and unbiased estimate is used for generalization error. Overfitting is a common issue in most of classifiers and regressors that decreases their generalization ability severely, especially when modeling high-dimensional data [20–26]. On contrast, the nature of randomness makes RF the trait of anti-overfitting for data with numerous variables. Dimensionality reduction is not necessary. Trees in RF are built independently with no branch pruning and easily parallelized, which further let RF be a simple and high-efficient modeling strategy. Additionally, RF gives useful internal estimates of error, importance and correlation of variables. These advantages make RF a desirable tool for predicting the attributes and categories of unknown samples [27–32]. In this study, we explore the potential of RF for modeling GC-MS metabolomics data of clinical urine samples for identifying six types of IMDs and normal ones simultaneously. The six types are methylmalonic acidemia (MMA), glutaric aciduria type I (GA I), propionic acidemia (PA), citrin deficiency (CD), isovaleric aciduria (IVA), and multiple acyl-CoA dehydrogenase deficiency (MADD). GC-MS metabolomics data are acquired for 447 urine samples from the newborns with different types of IMDs and 84 normal samples. Besides the RF algorithm, the artificial neural network (ANN) and support vector machine (SVM) are also used to modeling the same GC-MS data for comparison. The results have demonstrated that the proposed strategy affords highly sensitive and specific classification of all types of IMDs with high precision and accuracy, enabling high-efficient, robust and reliable identification of multiple IMDs for clinical diagnosis.

2. Experimental procedures

2.1. Equipment and reagents

Data were acquired on a gas chromatograph tandem a quadrupole mass spectrometer (GC-MS-QP2010, Shimadzu, Japan) equipped with an auto-sampler (GL 221–34618), an open split injector and helium as carrier gas.

The derivatizing reagents, BSTFA +1% TMCS (*N, O*-bis (trimethylsilyl) trifluoroacetamide with 1% trimethylchlorosilane) were purchased from ANPEL (Shanghai, China). Urease, margaric acid (MGA), sodium hydroxide and hydroxylammonium chloride were commercially obtained from Sigma-Aldrich company (St. Louis, MO, USA). Ethyl acetate, hydrochloric acid and picric acid were purchased from ANPEL (Shanghai, China), KAIXIN (Hunan, China) and Xiya (Chengdu, China), respectively. The solution of margaric acid (0.5 mg/mL) was used as internal standard, whose solvent was ethyl acetate.

2.2. Sample collection

The urine samples provided by Shenzhen Aone Medical Laboratory Co. Ltd. (Shenzhen, China) were collected from 447 patients (born within 28 days) diagnosed using genetic tests and 84 matched healthy controls. The positive samples were further divided into six groups, the MMA group of 257 instances, the GA I group of 30, the PA group of 40, the CD group of 50, the IVA group of 25 and the MADD group of 45. The collected urine samples were stored at −20 °C without the addition of any other reagents.

2.3. Sample preparation

The concentration of urinary creatinine was determined by picric acid method, and the volume of urine injected into GC-MS was adjusted to the same amount of creatinine in order to avoid measurement errors [33]. We added 20 μL of urease to the urine sample containing 0.2 mg creatinine. Next the sample was incubated at 37 °C for 30 min and then mixed with 0.02 mg internal standard solution. Then the mixture was diluted with distilled water to adjust the final volume to 2 mL. We added 1 mL of hydroxylammonium chloride (5%) and 400 μL of sodium hydroxide (20%) to the mixture, then it was left at room temperature for 60 min. The pH of the mixture was adjusted to 2–4 by adding 550 μL HCl (37%) and the solution was then shaken for about 3 min in a vortex mixer. Organic acids in the urine sample were extracted twice by adding 3 mL of ethyl acetate and fully mixed in a vortex mixer. The mixture was centrifuged for 5 min (4000 r/min). Two organic layers after centrifuging were transferred to a clean centrifuge tube and evaporated and dried under nitrogen at 60 °C. Finally, 100 μL of BSTFA: TMCS (99:1 in volume) was added to the tube and incubated at 70 °C for 30 min. After that, 1 μL of the final mixture was injected into the GC-MS apparatus.

2.4. Gas chromatography-mass spectrometry conditions

Chromatographic separation was performed on Shimadzu quadrupole mass spectrometer with a deactivated fused silica capillary column (Agilent DB-5, 30 m × 0.25 mm × 1 μm). The temperature was programmed from 100 °C to 280 °C at a rate of 4 °C/min. The temperatures of the injector, the transfer line and the ion source were set to be 280 °C, 280 °C and 200 °C. The mass spectrometer was performed under electron impact (EI) mode with ionization energy of 70 eV. And it was operated in a full scan mode with *m/z* range from 50 to 500 at the rate of 1000 Da/s.

2.5. Data preprocessing and statistical analysis

Generally, preprocessing of the raw data acquired from GC-MS is in order to provide a suitable format for later data analysis. Here, all GC-MS raw data of urine were processed using GC-MS solution software and the mass spectral database including mass spectral and retention index data of the metabolites (Shimadzu, Japan).

In this study, the concentrations of 132 metabolites in urine sample were identified. They were taken for the main endogenous metabolites closely related to about 40 kinds of organic acidurias. The contents of urinary metabolites were measured using internal standard method. The peak areas of the metabolites to be measured were compared with that of the internal standard to calculate their concentrations. Then the data matrix, including 257 MMA patients, 30 GA I patients, 40 PA patients, 50 CD patients, 25 IVA patients, 45 MADD patients and 84 healthy controls, was obtained. Within this matrix, the columns and rows indicated concentrations and samples, respectively. The matrix was used as the response matrix for RF modeling.

The entire data set was divided into two parts, a training set containing 265 samples and a test set containing 266 samples. The training set were composed of 128 samples taken randomly from the MMA group, 15 from the GA I group, 20 from the PA group, 25 from the CD group, 12 from the IVA group, 23 the MADD group and 42 from the control group. The remaining samples made up the test set.

3. Theory and algorithm

3.1. Random forest

The RF algorithm proposed by Breiman in 2001 is a superior algorithm for solving regression or multi-classification problems [34]. RF relies on two machine learning strategies, bagging and random variable selection [35]. Bagging is developed based on bootstrap sampling and has become the representative of parallel ensemble learning methods. Given a training set containing *m* samples, bootstrap sampling is performed for *m* times to give a sampling set containing *m* samples. Some samples in the training set may repeatedly present in the sampling set and some may not present after *m* times of sampling. In this manner, *T* sampling sets that each contains *m* training samples are generated. Learning machines can be trained based on each sampling set and then all the *T* learning machines are combined. RF is a variation of Bagging, composed a series of classification and regression trees (CARTs) as classifiers or regressors. The independent random vectors with identical distribution used in each classifier determine the growth of the decision tree. The majority vote of all the trees determines the output. In the training process, RF uses bootstrap sampling to generate multiple sub-training sets from the input training set and train multiple decision trees to improve the performance of the model. The constructive process of a decision tree is depicted in Fig. 1. In this process, each variable is selected from the subspace composed by a random combination of variables. For the decision tree grown on the sub-training set, out-of-bag (OOB) samples are used as the test set of the tree. with the increasing number of decision trees, RF gives an unbiased estimate of test set error according to the OOB data. In addition, OOB data can also be used to assess the importance of variables. OOB error and variable importance are two chief parameters of decision tree. The CART uses Gini index to evaluate variable importance to obtain the optimal variables and determine the optimal binary cut point of the variable. For the problem of multi-classification, assuming there are *K* categories and the probability of the *k*-th category is p_k, the Gini index expression of the distribution of probability is:

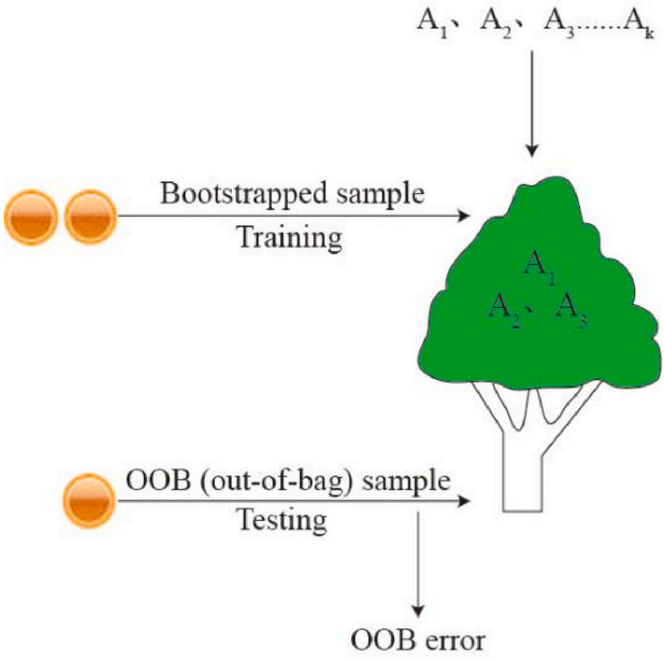


Fig. 1. The construction process of a decision tree.

$$Gini(p) = \sum_{k=1}^{K} pk(1-pk) = 1 - \sum_{k=1}^{K} pk^2 \tag{1}$$

For a given sample set *D*, its Gini index is:

$$Gini(D) = 1 - \sum_{k=1}^{K} \left(\frac{|Ck|}{|D|}\right)^2 \tag{2}$$

Here, C_k is the subset of samples belonging to the *k*-th class in *D*.

There are mainly three characteristics of RF: (1) The subset of training samples is randomly selected; (2) The subset of variables is randomly selected; (3) All the decision trees are left to grow without pruning. The implementation of random forest algorithm is summarized as follows (as shown in Fig. 2):

Step 1. Selection of the sample set

Since the sample numbers of the six types of IMDs were distributed unevenly, we optimized the sampling strategy in RF to avoid poor generalization error due to uneven distribution of samples. We performed bootstrap sampling on the training sets of six types of IMDs and healthy samples one by one. The bootstrap sampling was performed a total of n_{tree} ($n_{tree} = 200$) times that is the number of classification trees. The data after bootstrap sampling of each time of all types were collected to form a bootstrap training data set. Therefore, 200 training data sets were formed to grow 200 classification trees. The final training set contained two thirds of the original data, and the other containing one third of original data were employed as OOB samples. The OOB samples were then used to conduct internal validation of RF model.

Step 2. The generation of classification trees

For each bootstrap data set, the algorithm grew a classification tree which was unpruned and modified as follows: At each node of the classification tree, m_{try} variables were randomly selected and the best splits among these variables were determined instead of choosing the best splits among all the variables. Bagging was considered as a special case of random forest when $m_{try} = n$ (the total number of variables, *n* was 132 in this study). Generally, m_{try} was a positive integer between 1 and *n*. The default value of m_{try} was the square root of *n*, so we set m_{try} to be 12 in this study.

Step 3. Combination of the trees

Since the n_{tree} classification trees were constructed based on random combination of variables and samples, these trees were independent from each other. Therefore, the trees were equal on their importance. When they were combined, their weights were not need to be considered, or it could be considered having the same weight for every tree. The final output was determined by the votes of the trees. Then the sample could be predicted by aggregating the prediction results of the n_{tree} classification trees.

3.2. Evaluation of the model

The performance of the established classification model can be evaluated using the following indicators [36,37],

$$Specificity = \frac{TN}{N} \tag{3}$$

$$Sensitivity = \text{Recall} = \frac{TP}{P} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5}$$

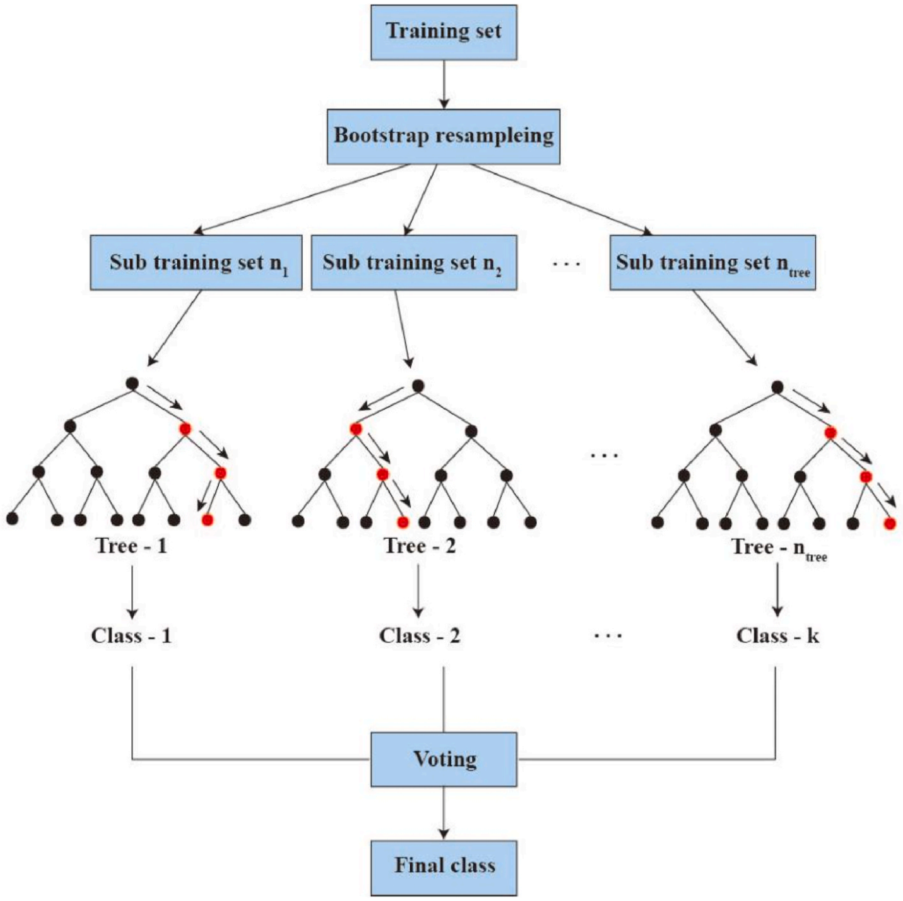


Fig. 2. The scheme of the processing of RF: starting from the original input training set (on the top), generating n_{tree} random sub-training sets (by bootstrap resampling) and training corresponding decision trees and voting, outputting the final class of samples.

$$Accuracy = \frac{TP + TN}{P + N} \tag{6}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{7}$$

where P is the number of positive samples, N is the number of negative samples, TP is the number of true positive samples, TN is the number of true negative samples, FP is the number of false positive samples, FN is the number of false negative samples. The *specificity* measures the ability to identify negative samples. The *sensitivity* is equal to the recall, measuring the ability to identify the positive samples. The *precision* refers to the proportion of true positive samples that identified accurately by the model. The *accuracy* is the discrimination accuracy of the model for both the positive and negative samples. The matthews correlation coefficient (*MCC*) is used to assess the predictive ability of classification models. In this study, although RF performs as a multi-classification strategy, one type of IMDs is taken as a positive sample and the rest are taken as negative samples while evaluating the performance of the model on classifying each category.

4. Results and discussion

4.1. GC-MS metabolomics data

Typical total ion current (TIC) chromatograms of urinary metabolites for healthy sample and six disease types of IMDs are shown in Fig. 3. The peaks of urinary metabolites distributed across a wide range of retention times from 5 min to 60 min. The abundant peaks suggested GC-MS an effective technology for the assays of human urinary metabolites. The intensities of some peaks were different for varying types of IMDs, which made the patterns of TICs not the same. However, no obvious features could be found for accurately identifying these six types of IMDs and healthy ones from each other, due to the complex relationships between the detectable metabolites. The results of traditional statistical analysis also did not reveal obvious evidences for multi-classification. To guarantee a high accuracy for classification, a more effective modeling technique that could extract and utilize the correlation of variables of GC-MS data was required.

4.2. Determination of the number of trees in RF

Considering the ability of RF in estimating the importance and correlation of variables of high-dimensional data, we tried RF for modeling GC-MS metabolomics data of IMDs. In the process of training the RF model, we first tried to determine the number of classification trees of RF. We estimated the trend of out-of-bag (OOB) classification errors with growing number of the trees built for RF reaching up to 200. As shown in Fig. 4, with the increasing numbers of trees, the OOB classification errors rapidly reduced in the beginning, and then entering a phase of gentle change. When the number of the grown trees was more than 80, the OOB classification errors reached to a minimum without any change until the tree number of 200. It implied the increasing number of trees more than 80 would not help improve the OOB classification error of the RF model. In this study, we set 100 as the optimized number of trees for establishing RF.

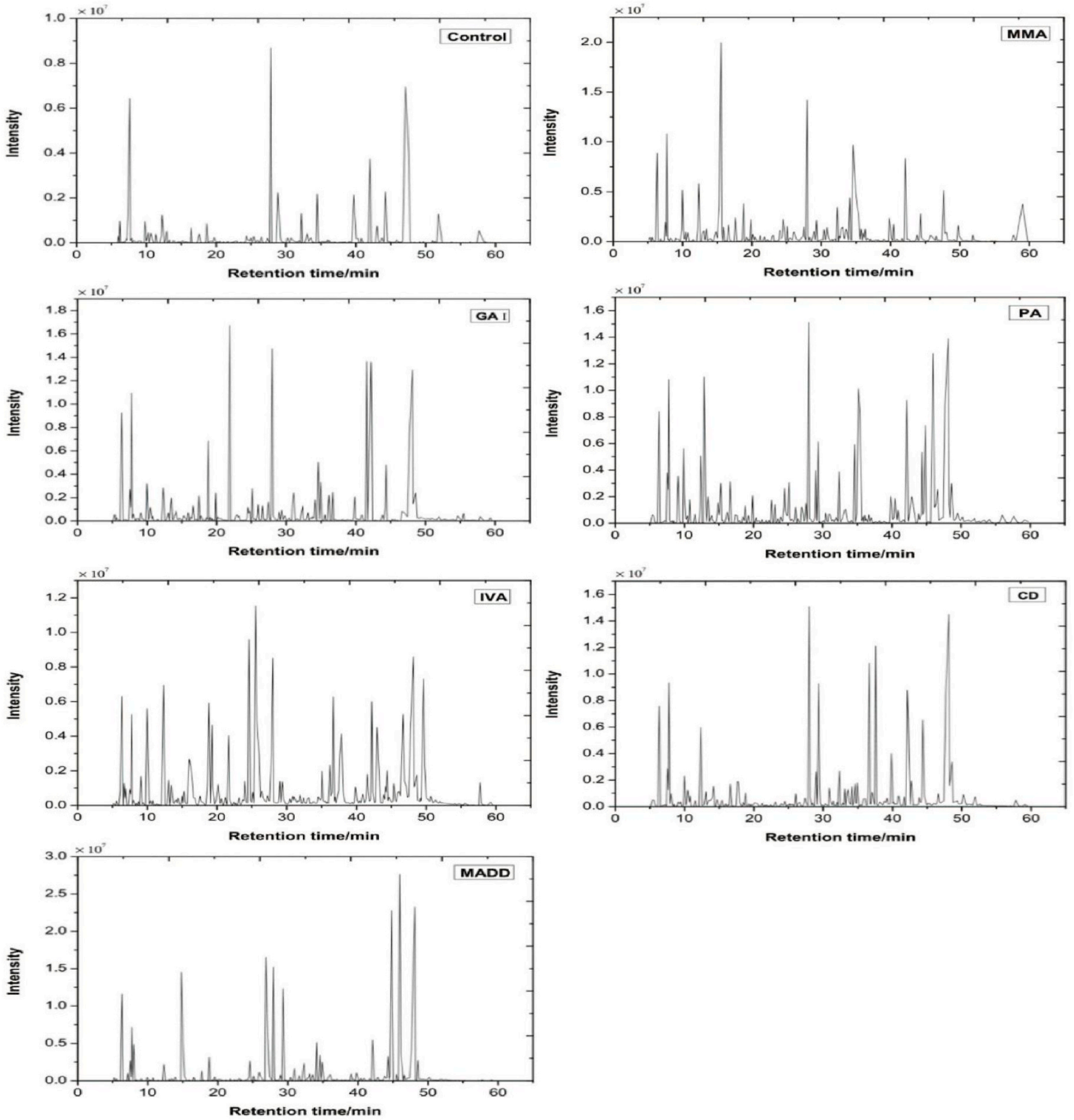


Fig. 3. Total ion current (TIC) chromatograms of urinary metabolites for healthy sample and six disease types of IMDs (MMA, GA I, PA, CD, IVA and MADD).

4.3. *Confusion matrix*

Confusion matrix was a visualization tool, especially for supervised learning. We used it to display the classification performance of a RF model on its test set (Fig. 5). Each row of the confusion matrix represented a prediction category, the total number of each row indicated the number of samples predicted as the corresponding category. Each column represented the true attribution of categories of the samples. The sum of each column is the total number of samples that actually belonging to the corresponding category. The green boxes on the diagonal of the confusion matrix indicated the prediction performance of the RF model for each category. For the number in each green box, the closer of its value to the sum of the numbers in its corresponding column, the better the prediction performance of the category corresponding to the column. It could be used as a measurement of the consistency between the prediction category and the actual one of the samples in each category. The percentage in each green box exhibited the portion of accurately predicted sample from a corresponding category to the total number of predicted samples from all categories. Otherwise, the red boxes indicated the inconsistency between the prediction categories and the actual ones of the samples. Therefore, the smaller values of the number and the percentage in each red box was, the better prediction performance of the RF model was each category of IMDs. As shown in the first column of the confusion matrix, 127 MMA samples were correctly identified out of 129 MMA samples in the test set. Two MMA samples were assigned into other categories, one for CD and the other for normal (healthy sample). Therefore, the sensitivity of the RF model in predicting MMA samples was estimated as 98.4%. Accordingly, the sensitivities of the model for predicting GA I, PA, CD, IVA, MADD and healthy samples were estimated to be 93.3%, 100%, 96.0%, 92.3%, 95.5% and 90.5%, respectively. Overall, 256 samples out of 266 samples in the test set were correctly assigned into their corresponding categories, which gave an overall sensitivity of 96.2% for the RF model. It thus suggested the desirable generalization performance of the RF model in multi-classification of GC-MS metabolomics data of IMDs. It was worth noting that, in practice for the GC-MS data with significant noise, a pretreatment procedure should be conduct in order to decrease the impact of the noise on the generalization ability of the RF model. Also,

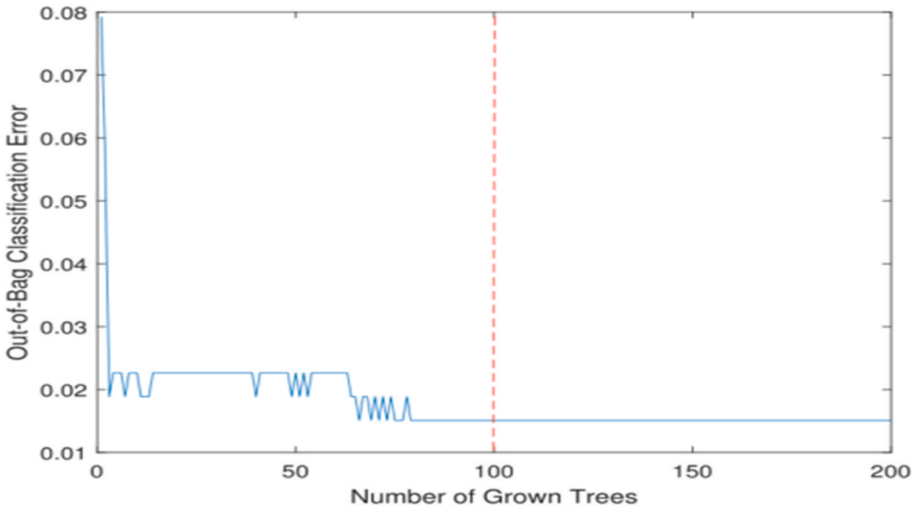
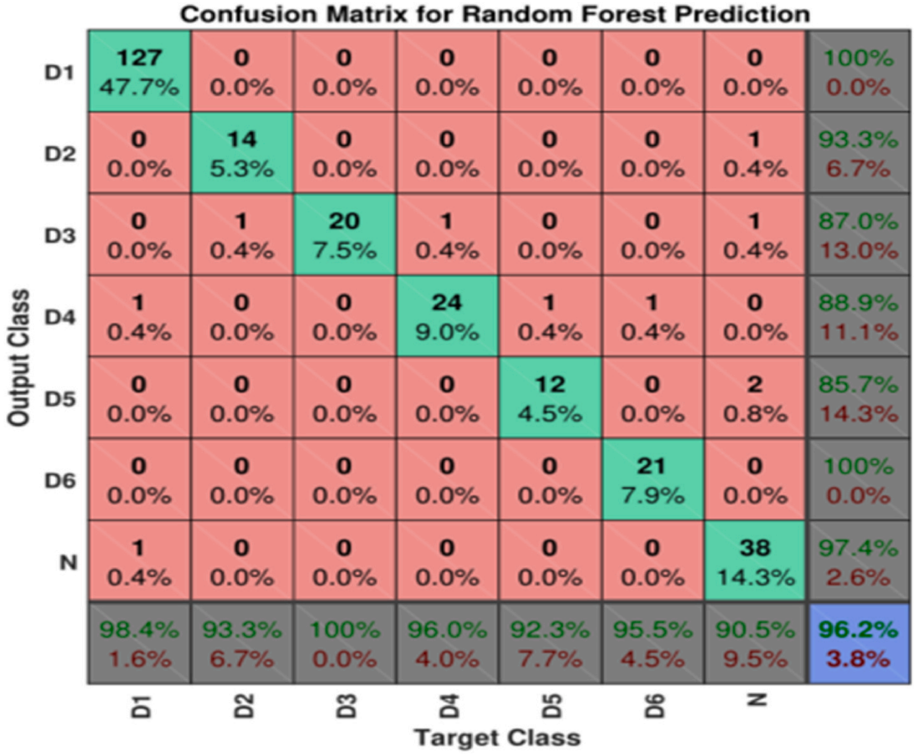


Fig. 4. The variation trend of OOB error along with the number of classification trees.


Confusion Matrix for Random Forest Prediction
Output Class
Target Class

	D1	D2	D3	D4	D5	D6	N	
D1	127 47.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
D2	0 0.0%	14 5.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.4%	93.3% 6.7%
D3	0 0.0%	1 0.4%	20 7.5%	1 0.4%	0 0.0%	0 0.0%	1 0.4%	87.0% 13.0%
D4	1 0.4%	0 0.0%	0 0.0%	24 9.0%	1 0.4%	1 0.4%	0 0.0%	88.9% 11.1%
D5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	12 4.5%	0 0.0%	2 0.8%	85.7% 14.3%
D6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	21 7.9%	0 0.0%	100% 0.0%
N	1 0.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	38 14.3%	97.4% 2.6%
	98.4% 1.6%	93.3% 6.7%	100% 0.0%	96.0% 4.0%	92.3% 7.7%	95.5% 4.5%	90.5% 9.5%	96.2% 3.8%

Fig. 5. The confusion matrix for RF model in prediction (D1, D2, D3, D4, D5, D6 and N represented the category of MMA, GA I, PA, CD, IVA, MADD and normal controls, respectively).

attention should be paid on the consistence between the training and predicting sets generated using different GC-MS instruments.

4.4. Comparison with other modeling methods

ANN and SVM were commonly used machine learning methods for building classification models. For comparison, we also used these two methods for modeling the GC-MS metabolomics data of IMDs. A set of indicators including *specificity, sensitivity, precision, accuracy,* and *MCC* for each category were calculated, as listed in Table 1. The ANN model was built with 10 nodes in the hidden layer and trained using back-propagation (BP) strategy. Sigmoid function was used as an activation function for the artificial neurons. To reduce the risk of ANN overfitting the training set, 15% of samples in the training set were randomly picked up and used as a calibration set for optimizing the network. The SVM model was built with sigmoid function for the kernel transformation. The results revealed that the ANN and SVM models both had acceptable specificity on identifying the samples those not belonging to corresponding categories, but showed poor sensitivity on identifying their true categories. Although the ANN model gave 94.74% of sensitivity for MMA and 100% for CD, that for the rest categories was lower than 80% and as low as 20% for IVA. The SVM model gave the best sensitivity of 92.19% for MMA, then 80% for CD, but also showed a poor sensitivity of 23.08% for IVA. Accordingly, the ANN and SVM models both gave the poorest precision to IVA as low as 25% and 30%, respectively. The highest *MCC* obtained using the ANN model was 1 for

Table 1
Results of ANN, SVM and RF for multi-classification of IMDs using GC-MS metabolomics data.

Method	Category	*Specificity* (%)	*Sensitivity* (%)	*Precision* (%)	*Accuracy* (%)	*MCC*
ANN	MMA	83.33	94.74	100.00	88.75	0.7819
	GAI	100.00	80.00	100.00	98.75	0.8885
	PA	98.68	50.00	66.67	96.25	0.5585
	CD	100.00	100.00	100.00	100.00	1
	IVA	96.00	20.00	25.00	97.33	0.1777
	MADD	96.10	67.67	40.00	95.00	0.4927
	Control	92.19	43.75	58.33	82.50	0.4026
SVM	MMA	77.54	92.19	79.19	84.59	0.7018
	GAI	100.0	60.00	100.00	97.74	0.7655
	PA	98.78	70.00	82.35	96.62	0.7415
	CD	98.76	80.00	86.96	96.99	0.8177
	IVA	97.23	23.08	30.00	93.61	0.2302
	MADD	99.59	34.78	88.89	93.98	0.5343
	Control	89.29	59.52	51.02	84.59	0.4591
RF	MMA	100.00	98.45	100.00	99.25	0.9851
	GAI	99.60	93.33	93.33	99.25	0.9293
	PA	98.78	100.00	86.96	98.87	0.9268
	CD	98.76	96.00	88.89	98.50	0.9156
	IVA	99.21	92.31	85.71	99.25	0.8837
	MADD	100.00	95.45	100.00	99.62	0.9750
	Control	99.55	90.48	97.44	98.12	0.9281

CD, but for the rest categories the *MCCs* were lower than 0.8885 with the poorest performance of just 0.1777 for IVA. The SVM model also gave the highest *MCC* of 0.8177 for CD and the lowest *MCC* of 0.2302 for IVA. Overall, the ANN and SVM models both showed unsatisfied performance on modeling the current GC-MS metabolomics data of IMDs.

For the model of RF, it had the specificity for each category ranging from 98.76% to 100%, and the sensitivity was equally higher than 90.48% and reaching to 100% for PA. The precision of the RF model was 100% for MMA and MADD, and the lowest was 85.71% for IVA. The accuracy for each category was higher than 98.12% and the *MCC*s were ranging from 0.8837 to 0.9851. The classification performance on each category had been greatly improved compared with those of the ANN and SVM models, suggesting the better generalization ability of the RF model than the ANN and SVM models. These results implied the promise of the RF strategy for robust and reliable classification of multiple IMDs based on GC-MS metabolomics data for clinical diagnosis.

5. Conclusion

We proposed a strategy for identifying multiple IMDs by combining the GC-MS technique for metabolomics data acquisition with RF algorithm for multi-classification. GC-MS is a cost- and time-effective technique for acquiring complex metabolomics data of urine samples. With the nature of anti-overfitting, the RF enabled a simple, high-effective and robust approach for modeling the high-dimensional GC-MS data. In total, 531 urine samples of newborns were collected and GC-MS analysis was performed to acquire the metabolomics data for building the RF model. The results revealed the RF model had quite desirable performance on multi-classification of IMDs with high specificity, sensitivity, precision, accuracy, and *MCC*s for all six types of IMDs and normal samples. It, therefore, afforded the proposed strategy of great promise in reliable and effective identification of multiple IMDs to serve clinical diagnosis.

Author contribution

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding sources

This work was supported by the National Natural Science Foundation of China (Grants 21874040 and 21521063), Shenzhen Science and Technology Innovations Committee (JCYJ2018050718342).

Notes

The authors declare no competing financial interest.

Credit author statement

Nan Chen: Conceptualization, Methodology, Software, Writing - review & editing. **Hai-Bo Wang**: Software, Validation, Writing - original draft. **Ben-Qing Wu**: Data curation, Software, Investigation. **Jian-Hui Jiang**: Conceptualization, Methodology, Project administration. **Jiang-Tao Yang**: Funding acquisition, Data acquisition. **Li-Juan Tang**: Conceptualization, Formal analysis. **Hong-Qin He**: Samples and data acquisition. **Dan-Dan Linghu**: Samples acquisition, Data curation.

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.

Acknowledgment

This study was financially supported by the National Natural Science Foundation of China (Grants 21874040, and 21521063), Shenzhen Science and Technology Innovations Committee (JCYJ2018050718342).

References

[1] M.J. Gambello, H. Li, Current strategies for the treatment of inborn errors of metabolism, J. Genet. Genomics. 45 (2018) 61–70, https://doi.org/10.1016/j.jgg.2018.02.001.

[2] A. Bower, A. Imbard, J.-F. Benoist, S. Pichard, O. Rigal, O. Baud, M. Schiff, Diagnostic contribution of metabolic workup for neonatal inherited metabolic disorders in the absence of expanded newborn screening, Sci. Rep. 9 (2019) 14098, https://doi.org/10.1038/s41598-019-50518-0.

[3] J.V. Leonard, A.A.M. Morris, Inborn errors of metabolism around time of birth, Lancet 356 (2000) 583–587, https://doi.org/10.1016/s0140-6736(00)02591-5.

[4] T. Fukao, K. Nakamura, Advances in inborn errors of metabolism, J. Hum. Genet. 64 (2019) 65, https://doi.org/10.1038/s10038-018-0535-7.

[5] M.H. Hampe, S.N. Panaskar, A.A. Yadav, P.W. Ingale, Gas chromatography/mass spectrometry-based urine metabolome study in children for inborn errors of metabolism: an Indian experience, Clin. Biochem. 50 (2017) 121–126, https://doi.org/10.1016/j.clinbiochem.2016.10.015.

[6] B.L. Therrel, C.D. Padilla, J.G. Loeber, I. Kneisser, A. Saadallah, G.J.C. Borrajo, J. Adams, Current status of newborn screening worldwide: 2015, Semin. Perinatol. 39 (2015) 171–187, https://doi.org/10.1053/j.semperi.2015.03.002.

[7] R. Guthrie, The origin of newborn screening, Screening 1 (1992) 5–15, https://doi.org/10.1016/0925-6164(92)90025-Z.

[8] R. Guthrie, The PKU Story, Hope Publishing House, Pasadena, CA, 1997, https://doi.org/10.1086/301960.

[9] R. Guthrie, A. Susi, A simple phenylalanine method for detecting phenylketonuria in large populations of newborn infants, Pediatrics 32 (1963) 318–343. https://pediatrics.aappublications.org/content/102/Supplement_1/236.short.

[10] H. Lemonde, M. Cleary, A. Chakrapani, Newborn screening for inborn errors of metabolism, J. Paediatr. Child Health 25 (2014) 103–107, https://doi.org/10.1016/j.paed.2014.10.010.

[11] G. Kaur, K. Thakur, S. Kataria, T.R. Singh, B.S. Chavan, G. Kaur, R. Atwal, Current and future perspective of newborn screening: an Indian scenario, J. Pediatr. Endocrinol. Metab. 29 (2016) 5–13, https://doi.org/10.1515/jpem-2015-0009.

[12] D. Ombrone, E. Giocaliere, G. Forni, S. Malvagia, G. la Marca, Expanded newborn screening by mass spectrometry: new tests, future perspectives, Mass Spectrom. Rev. 35 (2016) 71–84, https://doi.org/10.1002/mas.21463.

[13] M.Z. Peng, X.F. Fang, Y.L. Huang, Y.N. Cai, C.L. Liang, R.Z. Lin, L. Liu, Separation and identification of underivatized plasma acylcarnitine isomers using liquid chromatography-tandem mass spectrometry for the differential diagnosis of organic acidemias and fatty acid oxidation defects, J. Chromatogr. A 1319 (2013) 97–106, https://doi.org/10.1016/j.chroma.2013.10.036.

[14] T. Kuhara, Gas chromatographic-mass spectrometric urinary metabolome analysis to study mutations of inborn errors of metabolism, Mass Spectrom. Rev. 24 (2005) 814–827, https://doi.org/10.1002/mas.20038.

[15] D. Hori, Y. Hasegawa, M. Kimura, Y.L. Yang, I.C. Verma, S. Yamaguchi, Clinical onset and prognosis of Asian children with organic acidemias, as detected by analysis of urinary organic acids using GC/MS, instead of mass screening, Brain Dev. 27 (2005) 39–45, https://doi.org/10.1016/j.braindev.2004.04.004.

[16] D. Ombrone, E. Giocaliere, G. Forni, S. Malvagia, G. la Marca, Expanded newborn screening by mass spectrometry: new tests, future perspectives, Mass Spectrom. Rev. 35 (2016) 71–84, https://doi.org/10.1002/mas.21463.

[17] L.Z. Yi, N.P. Dong, Y.H. Yun, B.C. Deng, D.B. Ren, S. Liu, Y.Z. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: a review, Anal. Chim. Acta 914 (2016) 17–34, https://doi.org/10.1016/j.aca.2016.02.001.

[18] K.H. Liland, Multivariate methods in metabolomics-from pre-processing to dimension reduction and statistical analysis, Trac-trend. Anal. Chem. 30 (2011) 827–841, https://doi.org/10.1016/j.trac.2011.02.007.

[19] M.H. Hampe, P. Ingale, S.N. Panaskar, A.A. Yadav, Two tier analysis of organic acid disorders: a comprehensive approach for newborn screening, Int. J. Biomed. Adv. Res. 6 (2015) 84–90, https://doi.org/10.7439/ijbar.v6i2.1491.

[20] J. Bin, F.F. Ai, W. Fan, J.H. Zhou, Y.H. Yun, Y.Z. Liang, A modified random forest approach to improve multi-class classification performance of tobacco leaf grades coupled with NIR spectroscopy, RSC Adv. 6 (2016) 30353–30361, https://doi.org/10.1039/c5ra25052h.

[21] T. Sun, L. Jiao, F. Liu, S. Wang, J. Feng, Selective multiple kernel learning for classification with ensemble strategy, Pattern Recogn. 46 (2013) 3081–3090, https://doi.org/10.1016/j.patcog.2013.04.003.

[22] Q. Yang, S.S. Lin, J.T. Yang, L.J. Tang, R.Q. Yu, Detection of inborn errors of metabolism utilizing GC-MS urinary metabolomics coupled with a modified orthogonal partial least squares discriminant analysis, Talanta 165 (2017) 545–552, https://doi.org/10.1016/j.talanta.2017.01.018.

[23] K.P. Hsu, S.H. Hsieh, S.L. Hsieh, P.H. Cheng, Y.C. Weng, J.H. Wu, F.P. Lai, A newborn screening system based on service-oriented architecture embedded support vector machine, J. Med. Syst. 34 (2010) 899–907, https://doi.org/10.1007/s10916-009-9305-6.

[24] W.H. Chen, S.L. Hsieh, K.P. Hsu, H.P. Chen, X.Y. Su, Y.J. Tseng, Y.H. Chien, W. L. Hwu, F.P. Lai, Web-based newborn screening system for metabolic diseases: machine learning versus clinicians, J. Med. Internet Res. 15 (2013) e98, https://doi.org/10.2196/jmir.2495.

[25] Q. Yang, L. Xu, L.J. Tang, J.T. Yang, B.Q. Wu, N. Chen, J.H. Jiang, R.Q. Yu, Simultaneous detection of multiple inherited metabolic diseases using GC-MS urinary metabolomics by chemometrics multi-class classification strategies, Talanta 186 (2018) 489–496, https://doi.org/10.1016/j.talanta.2018.04.081.

[26] Q. Yang, L. Tan, B.Q. Wu, G.L. Tian, L. Xu, J.T. Yang, J.H. Jiang, R.Q. Yu, Beyond one-against-all (OAA) and one-against-one (OAO): an exhaustive and parallel half-against-half (HAH) strategy for multi-class classification and applications to metabolomics, Talanta 204 (2020) 104–107, https://doi.org/10.1016/j.chemolab.2020.104107.

[27] A. Mascellani, G. Hoca, M. Babisz, P. Kloucek, J. Havlik, 1H NMR chemometric models for classification of Czech wine type and variety, Food Chem. 339 (2021) 127852, https://doi.org/10.1016/j.foodchem.2020.127852.

[28] P. Baral, M.A. Haq, Spatial prediction of permafrost occurrence in Sikkim Himalayas using logistic regression, random forests, support vector machines and neural networks, Geomorphology 371 (2020) 107331, https://doi.org/10.1016/j.geomorph.2020.107331.

[29] X.Y. Li, T. Geng, W.J. Shen, J.R. Zhang, Y.Z. Zhou, Quantifying the influencing factors and multi-factor interactions affecting cadmium accumulation in limestone-derived agricultural soil using random forest (RF) approach, Ecotoxicol. Environ. Saf. 209 (2021) 111773, https://doi.org/10.1016/j.ecoenv.2020.111773.

[30] E.S. Gokten, C. Uyulan, Prediction of the development of depression and post-traumatic stress disorder in sexually abused children using a random forest classifier, J. Affect. Disord. 279 (2021) 256–265, https://doi.org/10.1016/j.jad.2020.10.006.

[31] L. Dai, C.M.V. Goncalves, Z. Lin, J.H. Huang, H.M. Lu, L.Z. Yi, Y.Z. Liang, D. Wang, D. An, Exploring metabolic syndrome serum free fatty acid profiles based on GC-SIM-MS combined with random forests and canonical correlation analysis, Talanta 135 (2015) 108–114, https://doi.org/10.1016/j.talanta.2014.12.039.

[32] L. Lebanov, L. Tedone, A. Ghiasvand, B. Paull, Random Forests machine learning applied to gas chromatography-Mass spectrometry derived average mass spectrum data sets for classification and characterisation of essential oils, Talanta 208 (2020) 120471, https://doi.org/10.1016/j.talanta.2019.120471.

[33] P. Emond, S. Mavel, N. Aïdoud, L. Nadal-Desbarats, F. Montigny, F. Bonnet-Brilhault, C. Barthélémy, M. Merten, P. Sarda, F. Laumonnier, P. Vourc'h, H. Blasco, C.R. Andres, GC-MS-based urine metabolic profiling of autism spectrum disorders, Anal. Bioanal. Chem. 405 (2013) 5291–5300, https://doi.org/10.1007/s00216-013-6934-x.

[34] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32. https://link.springer.com/article/10.1023/A:1010933404324.

[35] J.H. Huang, L. Fu, B. Li, H.L. Xie, X.J. Zhang, Y.J. Chen, Y.H. Qin, Y.H. Wang, S. H. Zhang, H.Y. Huang, D.F. Liao, W. Wang, Distinguishing the serum metabolite profiles differences in breast cancer by gas chromatography mass spectrometry and random forest method, RSC Adv. 5 (2015) 58952–58958, https://doi.org/10.1039/c5ra10130a.

[36] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochem. Biophys. Acta. 405 (1975) 442–451, https://doi.org/10.1016/0005-2795(75)90109-9.

[37] Y. Zhang, L.J. Tang, H.Y. Zou, Q. Yang, X.L. Yu, J.H. Jiang, H.L. Wu, R.Q. Yu, Identifying protein arginine methylation sites using global features of protein sequence coupled with support vector machine optimized by particle swarm optimization algorithm, Chemometr. Intell. Lab. Syst. 146 (2015) 102–107, https://doi.org/10.1016/j.chemolab.2015.05.011.