Predictive Modeling for Metabolomics Data

Chapter *in* Methods in molecular biology (Clifton, N.J.) · January 2020

DOI: 10.1007/978-1-0716-0239-3_16

CITATIONS
38

READS
322

4 authors, including:

Debashis Ghosh
University of Colorado
434 PUBLICATIONS **29,650** CITATIONS
SEE PROFILE

Katerina Kechris
University of Colorado
339 PUBLICATIONS **4,794** CITATIONS
SEE PROFILE

Chapter 16

Predictive Modeling for Metabolomics Data

Tusharkanti Ghosh, Weiming Zhang, Debashis Ghosh, and Katerina Kechris

Abstract

In recent years, mass spectrometry (MS)-based metabolomics has been extensively applied to characterize biochemical mechanisms, and study physiological processes and phenotypic changes associated with disease. Metabolomics has also been important for identifying biomarkers of interest suitable for clinical diagnosis. For the purpose of predictive modeling, in this chapter, we will review various supervised learning algorithms such as random forest (RF), support vector machine (SVM), and partial least squares-discriminant analysis (PLS-DA). In addition, we will also review feature selection methods for identifying the best combination of metabolites for an accurate predictive model. We conclude with best practices for reproducibility by including internal and external replication, reporting metrics to assess performance, and providing guidelines to avoid overfitting and to deal with imbalanced classes. An analysis of an example data will illustrate the use of different machine learning methods and performance metrics.

Key words Metabolomics, Mass spectrometry, Supervised learning, Performance Metrics, Predictive Modeling

1 Introduction

In the past 20 years, there has been a dramatic increase in the development and use of high-throughput technologies for measuring various types of biological activity. Common examples include transcriptomics (the measurement of gene expression) and proteomics (the measurement of protein levels). The focus of this chapter is on metabolomics, which involves the measurement of small compounds, referred to here as metabolites, on a high-throughput basis. As products of activity at the protein level, metabolites represent an intermediate level between regulatory processes such as methylation and transcription, and the full spectrum of physiological and disease states. One appealing feature of metabolites is their ability to be used as clinical biomarkers, and for this reason, metabolomics has been extensively applied for finding biomarkers and

Shuzhao Li (ed.), *Computational Methods and Data Analysis for Metabolomics*, Methods in Molecular Biology, vol. 2104, https://doi.org/10.1007/978-1-0716-0239-3_16,

studying physiological processes and phenotypic changes associated with disease [1–4].

Metabolomics experiments fall into two categories: targeted and untargeted. Targeted metabolomics experiments measure ions from known biochemically annotated metabolites. By contrast, untargeted metabolomics experiments measure all possible ions within a predefined mass range and as a result may also include ions that do not map to known metabolites [5–8]. The main objective of metabolomics is to quantify and characterize the whole spectrum of metabolites. There are a variety of platforms by which metabolites can be measured. Examples include gas chromatography mass spectrometry (GC-MS) and liquid chromatography mass spectrometry (LC-MS) [9–11]. At a high level, these platforms input a sample, fragment it into ions, and separate them using physical properties in order to generate spectra for the sample. The fragmented ion spectra are then selected based on their physical properties (e.g., the retention time and the mass–charge ratio). In many instances, these properties can be used to map the ions to known metabolites.

Metabolomics data pose a variety of analytical challenges [12, 13]; thus, carefully constructed analytical pipelines need to be developed in order to preprocess and normalize the data. Once the data are normalized, one can proceed with various downstream analytical tasks, such as differential expression analysis, clustering, classification, network discovery, and visualization. In this chapter, we focus on the particular task of classification, which also goes by the name of prediction, supervised learning and biomarker discovery. We give an overview on some of the most commonly used methods for classification, along with an illustrative example using a dataset from our group. The structure of this chapter is as follows. In Subheading 2, we provide a short review of missing values and techniques for missing value imputation in metabolomics data. We then briefly describe the most commonly used supervised learning methods (Random Forest, Support Vector Machine, and Partial Least Square-Discriminant Analysis). In Subheading 3, we lay out a framework on fitting prediction models and their practical issues. This is followed by an illustrative example of data analysis and performance evaluation in Subheading 4, and the chapter ends with a short discussion in Subheading 5. In this chapter, we interchangeably use the term supervised learning, predictive modeling and machine learning.

2 Methods

2.1 Missing Values

Supervised learning methods require complete data; however, untargeted metabolomic data is prone to missing values, where the data matrix contains zeros in one or more entries. Some studies

have reported 20–30% missing values in datasets generated using untargeted MS [14, 15]. It is difficult to deduce whether a missing value is a genuine absence of a feature, a feature below the lower limit of detection of the machine, or the failure of the algorithms employed to identify real signals from the background. In practice, statisticians have defined three types of mechanisms that lead to missing values: missing at random (MAR), missing not at random (MNAR) and missing completely at random (MCAR) [16–18]. MCAR means that the missingness mechanism is completely random and depends neither on the observed data nor on the missing data. Scientifically plausible reasons that are compatible with missing completely at random include random errors or stochastic fluctuations of peak detection during the acquisition process of the raw data (incomplete derivations of signals). MAR means that the probability of a variable being missing is fully accounted for by other observed variables. Missing not at random (MNAR) means that the missingness mechanism depends on the unobserved values. If analysts believe MNAR to hold, there are unfortunately no ways to assess this assumption using observed data. A practical strategy is to collect as much covariate information as possible in order to make the MAR assumption plausible with the observed data.

There have been several attempts in the literature to deal with missing values for metabolomics data. For example, fillPeaks [19] in the XCMS software package has many missing value imputation tools available. A practical rule of thumb is to impute missing values by a small value or zero. This is problematic in that this leads to distortions of the distribution of missing variables and can cause the standard deviations to be underestimated [20]. Finally, Zhan et al. developed kernel-based approaches which explicitly modeled the missingness into a differential expression analysis [21]. Other imputation strategies include imputing missing values by zero, half of the minimum value or by the mean or median of observed values. More advanced methods use, random forest (RF) [20, 22], singular value decomposition (SVD) [23, 24], and *k*-nearest neighbors (kNN) [25]. The choice of these methods can influence the data analyses and inferences [14, 22, 26]. It is therefore extremely crucial to select the most suitable method for tackling missing values before moving forward with prediction. Recent work has compared performance of various missing value imputation methods [14, 25, 27, 28] on MS metabolomics data [20].

2.2 Classification Methods: An Early Look

It is important to note that what we now call supervised learning dates back to over 80 years ago, when Sir R. A. Fisher introduced the use of linear discriminant analysis (LDA) [29]. This was a generative model in which the features conditional on class label were modeled as a multivariate normal distribution with a mean vector that depended on group, and a common covariance matrix.

This was generalized to quadratic discriminant analysis, in which the covariance matrix also depends on the group. Linear discriminant analysis possessed two desirable properties:

1. Since the multivariate normal distribution is fully specified by the mean vector and covariance matrix, it is relatively simple to compute.
2. The classification rule from LDA is linear in the predictors and thus simple to interpret.

While this methodology is well established, there are two challenges with modern metabolomics data that make the utility of LDA less effective. First, in most situations, the number of metabolites being measured is greater than the sample size, which means that the covariance matrix will not be directly estimable from the observed data. Second, there is an increasing recognition that the linear classification rule might be too restrictive and that analysts should consider other nonlinear classifiers. This will motivate the classification tools we describe in Subheadings 2.3–2.5.

A second technique that dates back to the 1950s and has been used extensively in machine learning is Naïve Bayes [30]. In this framework, we assume the features are conditionally independent given the group label and model the likelihood ratio of the feature given the group label. Based on the product of these likelihood ratios, we are able to assign a new observation to a predicted group. The term "Naïve" comes from the fact that we assume that the features are statistically independent when, in fact, we know that they are not. That said, Naïve Bayes has been shown to be an effective tool in classification problems [31], and it can handle the situation when the number of metabolites measured is greater than the sample size.

2.3 Decision Tree

A Decision Tree (DT) is a supervised machine learning model, that outputs a hierarchical structure to classify subjects [32]. It is a nonlinear classifier which is mainly used for classifying nonlinearly separable data. The objective of a decision tree is to develop a model that predicts the value of a response variable based on several predictor variables. Figure 1 shows an example of a hypothetical DT, which divides the data into two categories based on two input variables. DT used in data mining can be classified into two groups:

- Classification tree: The predicted outcome is a categorical variable, representing two or more classes to which the observation belongs.
- Regression tree: The predicted value is a continuous variable.

DT is also known as Classification and Regression Trees (CART), which was first introduced in the machine learning literature [33]. The main difference between classification and regression trees is the criteria on which the split-point decision is made.

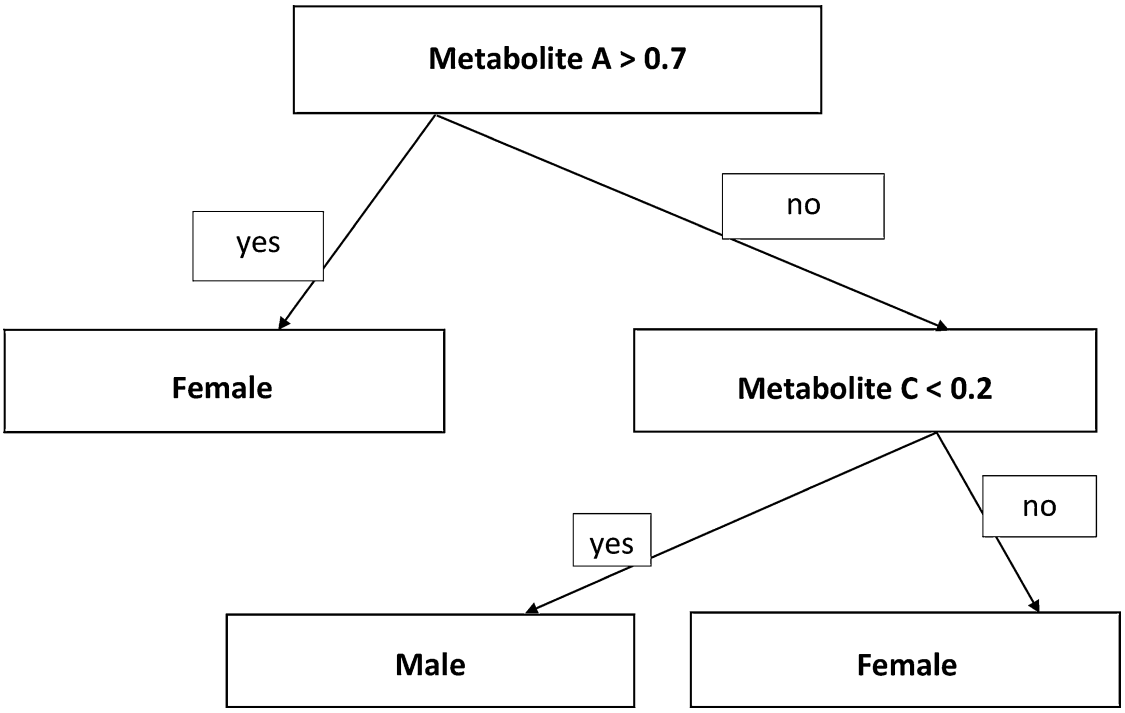

Fig. 1 A simple decision tree that splits the data into two gender groups based on two metabolites

2.4 Random Forest

A Random Forest (RF) is an extremely reliable classifier and robust to overfitting. It constructs an ensemble of DTs, which means an aggregation of tree-structured predictors [34]. In RF, each tree is independently constructed using a bootstrap sample of the original data (the "bagged sample"). This training data is used to build the classification model. The data that was not sampled using the bootstrap is referred to as the out-of-bag sample. Since these data were not used in model building, they can be used as a test data set, which can be used to evaluate classification accuracy in an unbiased manner, by calculating the "out-of-bag error" [35]. A measure of the variable importance of classification is also computed by considering the difference between the results from the original and randomly permuted versions of the data set. Cross-validation is not needed since RF is estimated from the bootstrap samples.

RF has become popular as a biomarker detection tool in various metabolomics studies [36, 37]. RF has the strength to deal with missing and data [34, 38] and overfitting issues [39, 40]. In addition, it can also tackle high-dimensional data sets without feature elimination as a requirement [41].

2.5 Support Vector Machines

Support Vector Machines (SVM) have been previously used in the analysis of several omics studies, particularly gene expression data [42–44]. A simple figure of an SVM is shown in Fig. 2. The main characteristics that define the concept of SVMs are (a) the criteria they use to categorize nonlinear relationships (b) the set of training

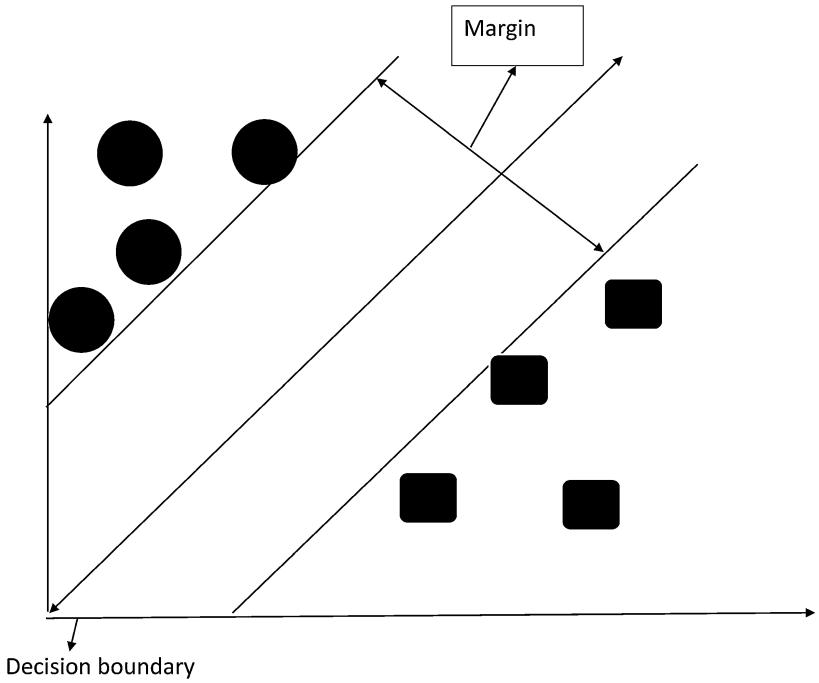

Fig. 2 A simple graphical representation of SVM

sets that are necessary to optimize the linear classifier; (c) the use of kernel machines to transform the variable into a higher order nonlinear space where linear separability holds; (d) utility in terms of performance and efficiency for high dimensional data sets.

A major drawback of SVM is its restrictions to binary classification problems. For example, it can only discriminate between two classes where the data points are categorized by two classes in n-dimensional space, where n corresponds to the number of metabolites in our context. A hyperplane is constructed that separates the data points from the two classes. The hyperplane coefficients are determined based on the variable (metabolite) importance for discriminating between two classes.

SVM can yield a hyperplane of p-1 dimension in p dimensional space. The main purpose of SVM is to optimize the largest margin. In practice, a separation often does not exist as the data points cannot always be linearly separated. In such nonlinear cases, a kernel substitution is adopted to map the data to a higher order dimension. The maximum-margin hyperplane was the original algorithm developed as a linear classifier [45]. An extension to create nonlinear classifiers was proposed by applying the kernel trick to maximum-margin hyperplanes [46]. The advantage of using the kernel trick is that it can substitute the linear kernel with other robust kernels, such as the Gaussian kernel [47]. Also in the family of nonlinear supervised learners are deep neural networks (DNN), which construct a nonlinear function from input variables to outcome variables using a combination of convolution filters and hidden layers [48].

2.6 Partial Least Squares-Discriminant Analysis

Partial least squares-discriminant analysis (PLS-DA) is a supervised technique widely used in metabolomics studies [49–52]. It is mainly constructed on the rotation of metabolite abundances in order to maximize the covariance between the independent variables (metabolite abundances) and the corresponding response variable (classes) in high-dimension by finding a linear subspace of the predictors [53]. PLS-DA is an extension of classical PLS regression which was implemented for solving linear equations and estimating parameters of interest. PLS-DA method has been extensively used in various metabolomics studies for disease classification and biomarker detection [50, 54–56]. Furthermore, PLS-DA can also be used for dimension reduction, and feature selection by ranking the loading vectors in decreasing order [52, 57, 58].

Orthogonal PLS (OPLS)-DA was developed as an improvement to PLS-DA in order to discriminate two or more classes of metabolites using multivariate data [59, 60]. The main advantage of OPLS-DA over PLS-DA is that a single component is used as a predictor where the other components constitute the orthogonal contrasts for analysis of variance, which are independent linear comparisons between the classes of a component.

Multilevel PLS-DA is another classification technique that can be used to classify multivariate data from crossover designed studies [61]. For example, each subject in a controlled experimentation setup undergo treatment in a random order [62]. Multilevel PLS-DA can be thought as a multivariate extension of the paired *t*-test [61].

3 Practical Issues in Fitting Prediction Models

3.1 Feature Selection

Feature Selection (FS) is an important step in successful data mining procedures [63], such as SVMs [64, 65] and Naïve Bayes [66], to enhance performance and reduce computational efficiency. However, FS is not a necessary criterion for some supervised algorithms, such as SVM due to its reliance on regularization, which is the process of adding information to prevent overfitting in order to enhance the predictive accuracy and interpretability of the supervised learning model. The purpose of feature selection is similar to model selection [67], which tries to find a compromise between high predictive accuracy and a model with few predictors. The insignificant input features in a supervised model may lead to overfitting. Hence, it is reasonable to ignore those input features with negligible or no effect on the output. For example, in the example later in this chapter, the objective is to infer the relationship between gender and their corresponding metabolite features. However, if the sample identifier or any other redundant column is included as one of the input features, it may cause overfitting. FS

is generally used as a preprocessing tool, in order to reduce the dimension of a data set by only selecting subsets of features (metabolites), on which a supervised learning is employed. Some well-known extensions of these FS techniques are Recursive Feature Elimination, L1 norm SVM [68], and Sequential Minimal Optimization (SMO) [69].

One of the most commonly used measure in FS is the Variable Importance Score (VIS), which evaluates features using a model-based approach [70] by ranking the features according to their relevance in a classification problem [71]. The main advantage of using VIS is that incorporates the correlation structure between the predictors (metabolite features) into the importance calculation.

3.2 Cross-Validation

The classification performance of supervised learners is crucial to determine their predictive power and accuracy. Generally, the validation procedures are implemented by assuming the model on a training set and then testing it on an independent set (validation data set). However, in practical situations, due to the relatively small number of samples and unavailability of an unbiased independent validation data set, cross-validation (CV) can be applied by splitting a data set into training and test sets. Using k-fold cross validation [36], the training data set is split into k subsets (folds) of almost equal size, that is, where k-1 training sets consist of x% of the data and the remaining (100-x)% data is contained in the kth test data set. Ideally, x% far exceeds (100-x)%, and x is usually chosen as 90, 80 or 70. Leave-One-Out-CV is a special case of CV, where k is equal to the total number of data points.

3.3 Metrics for Evaluation

There are several potential metrics by which one can evaluate a prediction model. The most common metric that is used in practice is the classification accuracy, meaning the proportion of predictions from the model that are correct based on the gold standard label. An alternative classification metric is given by the receiver operating characteristic (ROC) curve. Assume that we have two groups, disease and control and that higher values of the model correspond to a greater probability of having disease. We will let the model output be Υ and group label be D, where $D = 0$ means control and $D = 1$ means diseased. One can define the false positive rate based on a cutoff c by $\mathrm{FP}(c) = P(\Upsilon > c | D = 0)$. Similarly, the true positive rate is $\mathrm{TP}(c) = P(\Upsilon > c | D = 1)$. The true and false positive rates can then be summarized by the receiver operating characteristic (ROC) curve, which is a graphical presentation of $\mathrm{TP}(c)$,$\mathrm{FP}(c)$ for all possible cutoff values of c. The ROC curve shows the tradeoff between increasing true positive and false positive rates. Then, the area under the ROC curve (AUC) can be measured for the curve and is a summary based on how well the model can distinguish between two diagnostic groups (diseased/control). Other commonly used metrics are defined in terms of $\mathrm{TP}(c)$ and $\mathrm{FP}(c)$ as below:

Sensitivity (SENS): SENS = TP(c).

Specificity (SPEC): SPEC = 1 − FP(c).

Precision (PREC): TP(c)/(TP(c) + FP(c).

Recall (REC): REC=SENS = TP(c).

False Discovery Rate (FDR): FP(c)/(TP(c) + FP(c).

The predicted classes are conventionally computed based on the cut-off c (=50%) for the probabilities. However, the cutoff (threshold) value can be tuned to control the FDR depending on the problem setting in order to attain maximum predictive accuracy.

Calibration is another property that has been espoused for risk prediction models. Well-calibrated models are those in which the predicted risk matches the observed risk for individuals. The manner in which this is typically assessed is by comparing the risk predictions from the model to some nonparametric (i.e., non–model-based) estimate; the closer the predictions are, the better calibrated the model is. Calibration has been advocated in the risk prediction [72]. As a matter of course, nonparametric estimates of risk models require binning of covariates or categorization of predicted values in order to deal with the inherent sparsity that exists with using continuous covariates. One method of performing calibration, in the binary outcome setting, is to use the Hosmer–Lemeshow goodness of fit statistic [73]; smaller values of the statistic correspond with better calibrated models.

In the calibration setting, what is important is understanding the distribution of the predicted probabilities, or equivalently, the risk scores, from the fitted model. Calibration of the model then is equivalent to modeling the distribution of risk scores; a useful quantity for accomplishing this is the predictiveness curve [74].

3.4 Imbalanced Classes

In numerous data sets, there are unequal numbers of cases in each class. In this instance, the classifier is biased toward better performance of the larger (or majority) class, compared to the smaller (or minority) class. Often, the research question is much more focused on performance of discriminating the minority class from the majority class. But the size of the minority class may be limited by the difficulty, expense, or time of obtaining the rarer type of sample. This unequal distribution between classes of a data set is referred to as the imbalanced class problem [75].

In such cases, the main interest lies in the correct classification of the "minority class" [76]. Classes with fewer samples or no sample have a low prior probability and low error cost [77]. The relation between the distribution of samples in the training set and costs of misclassification can be controlled by setting a prior probability at each class.

Several methods have been discussed for tackling imbalanced data [78, 79], and two techniques which have been extensively applied in the last decade are resampling and cost-sensitive learning. In resampling, the approach is to either oversample the minority class or undersample the majority class. For example, the minority class can be oversampled by producing duplicates [80] or under sampling (removing samples) of the majority class [81, 82]. One major drawback of under sampling is that the majority class may lose some information, if a large part of majority class in a small training set is not considered. In cost-sensitive learning, the approach is to assign a cost misclassification of the minority class and minimize the overall cost function [83, 84]. Both the resampling and cost-sensitive learning approaches are considered to be more effective in terms of predictive accuracy than by using equal class prior constraints [85].

3.5 TRIPOD Guidelines

A recent scientific initiative has focused on developing more reproducible approaches to the building, evaluation and validation of prediction models. A document resulting from this effort that helps in this goal is the TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement, which provides recommendations for fair reporting of studies developing, validating, or updating a prediction model. It consists of a 22-item checklist detailing vital information that must be incorporated in a prediction model study report [86]. For our example analysis below, we provide supporting information to illustrate how our analysis satisfies the 22-item TRIPOD checklist (Appendix 1).

4 Illustrative Example

4.1 Data

For the predictive comparison of classifiers, we obtained a LC-MS metabolomics dataset from https://www.metabolomicsworkbench.org (Project ID: PR00038). The data were generated from subjects enrolled in the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease Gene study (COPDGene) [87, 88]. Plasma from 131 subjects was collected from the COPDGene study cohort and analyzed using untargeted LC-MS (C18+ and HILIC+) metabolomics. The lipid fraction of the human plasma collected from current and former smokers was analyzed using Time of Flight (ToF) liquid chromatograph (LC) (Agilent 6210 Series) and a Quadrupole ToF mass spectrometer (Agilent 6520) which yielded combined data on 2999 metabolite features. Data were annotated, normalized and preprocessed using the methods described in [87, 88].

COPD is an extremely heterogeneous disease comprising multiple phenotypes. The 131 subjects were either current or former

smokers with various chronic obstructive pulmonary disease (COPD) phenotypes including airflow obstruction, radiologic emphysema, and exacerbations. Within this set there were 56 males and 75 females. For additional information about the cohort, sample collection and data storage data generation, *see* [87].

4.2 Training and Test Sets

We split the data (131 samples) into 70% (93 samples) training and 30% (38 samples) test (evaluation) data. For the training data, we use fivefold CV, where we split the training data (93 training samples) into 5 different subsets (or fivefolds). We used the first four-folds to train the data and left the last (fifth) fold as holdout-test dataset. We then trained the algorithms against each of the folds and computed (average over fivefolds) the metrics for the training dataset. The test dataset ($n = 38$ samples) is used to provide an unbiased evaluation of the best model fit on the training dataset. The test dataset can be regarded as an external dataset which provides the gold standard used to evaluate the models, using ROC curves and other metrics for evaluation. For model validation, we predicted the performance of the test data using the trained models for all the three classifiers.

4.3 Feature Ranking and Variable Importance

In this section, we implemented different predictive models using metabolite abundances as the predictor variables and Gender (Male/Female) as the response based on the training dataset. We then computed the Variable Importance Score, which is a measure of feature relevance to gender for each metabolite (*see* Subheading 3.1). These scores are nonparametric in nature, and range between 0 and 100. They are subsequently used to rank all the features to the classification of our response variable, that is, Gender. Metabolites with high values are considered to more relevant features in classification problem.

In the dataset, the top five metabolites are detected as feature metabolites out of 2999 metabolites in the training set with fivefold Cross-Validation for three different classifiers (Fig. 3a–c). Among them, C39 H79 N7 O + 7.3314843, N-palmitoyl-D-sphingosyl-1-(2-aminoethyl)phosphonate, and C43 H86 N2 O2 are considered to be significant metabolite features based on RF and SVM classifiers. However, zeta-Carotene, unannotated metabolite (mass: 2520.6355 and retention time: 1.5409486), 5-Hydroxyisourate +4.668069, C13 H28 N2 O4, and Tyrosine∗ +2.3151746 are identified as good predictors based on PLS-DA.

4.4 Model Validation

In this section, we evaluated the performance of all the three classifiers based on the 30% test data of 38 samples using the trained models. Here, we present ROC curves for all the predictive models of the testing data used to compute the diagnostic potential of a classifier in this clinical metabolomics application. From the ROC

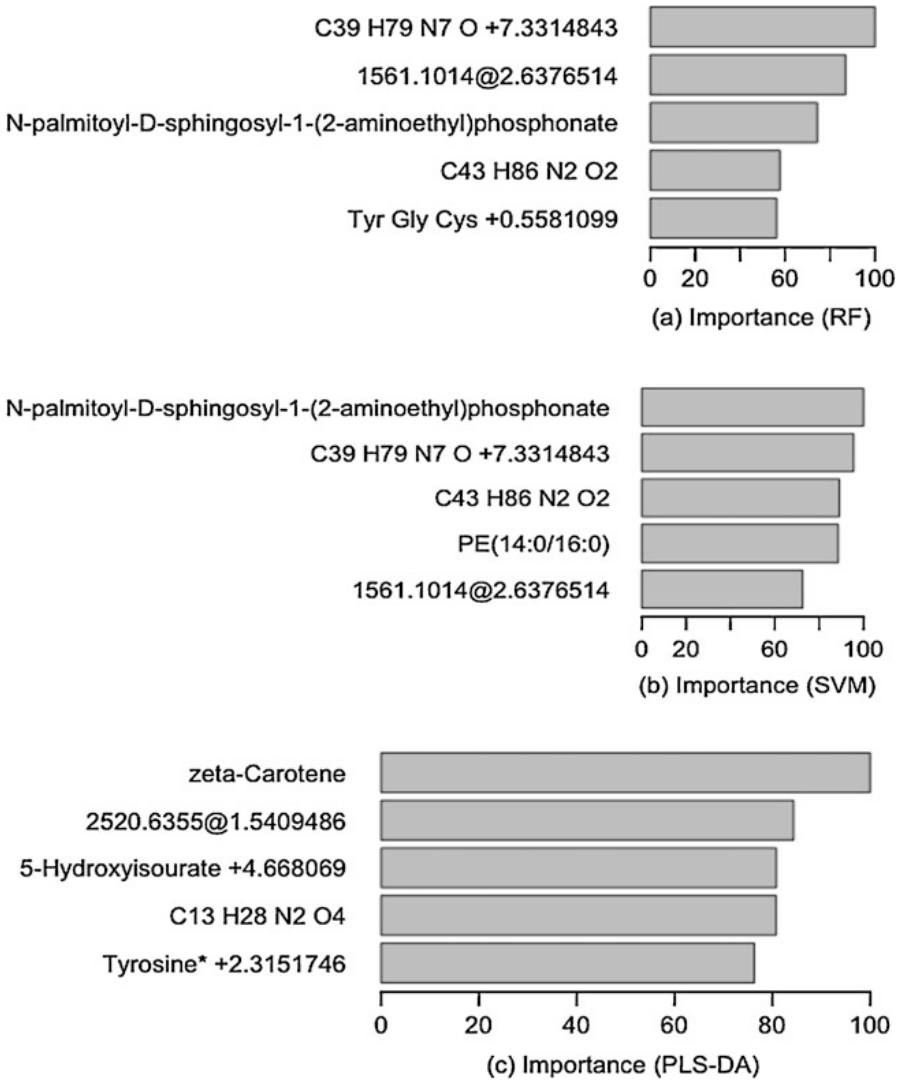

Fig. 3 Metabolite relevant feature ranking bar plots (top five metabolites) using Variable Important Scores ranging from 0 to 100. (**a**) Random Forest, (**b**) Support Vector Machine (SVM), and (**c**) Partial Least Square-Discriminant Analysis (PLS-DA) for the training dataset

curves, the three methods perform similarly (Fig. 4). Table 1 shows the performance metrics of the testing data evaluated for all the classifiers. In this testing dataset, we use AUC as our metric to choose the best performing classifier. Based on this metric RF has a small advantage over the other methods (0.87 versus 0.86), but with other metrics the other methods have a small advantage. In addition, we also computed the Variable Importance Score on the test dataset. The top five metabolites for all the three classifiers using the test dataset were exactly the same selected using the training dataset with fivefold CV in the previous section.

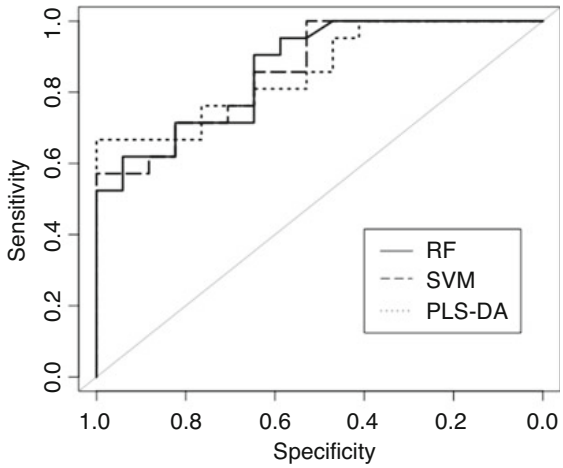

Fig. 4 ROC curves of the testing dataset obtained from three classification algorithms (RF, SVM, and PLS-DA)

Table 1
Metrics (area under curve (AUC), sensitivity (SENS), specificity (SPEC), precision (PREC), recall (REC)) to evaluate the performance of classification on testing dataset

Metrics/methods	AUC	SENS	SPEC	PREC	REC
RF	0.87	0.71	0.64	0.71	0.71
SVM	0.86	0.76	0.71	0.76	0.76
PLS-DA	0.86	0.81	0.65	0.74	0.81

5 Summary

Biomarker detection in the field of metabolomics is popular both in the context of prognostic and diagnostic studies. In this chapter, we discussed the most commonly used supervised learning algorithms, feature selection methods, and performance metrics, used in the downstream analyses of metabolomics studies. In addition, we also reported predictive accuracy of three classifiers on an example human plasma LC/MS test dataset to predict gender. Even though there were advantages of one method compared to the other depending on the metric, our results cannot be held as a comprehensive comparison of these methods, since different classifiers perform differently depending on the datasets. We encourage investigators to explore a variety of methods. For more detailed discussions of biomarker detection and predictive accuracy, *see* [89–92]. The R code for this chapter is posted at the supplemental website, https://metabolomics-data.github.io/. Appendix 2 lists selected open source tools that implement supervised learning algorithms.

TRIPOD Checklist for Predictive Modeling for Metabolomics Data

Section/topic	Item	Checklist item	Section
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted	See title
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions	See abstract
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models	Subheading 4.1
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both	Internal validation, Subheading 4.4

(continued)

Section/topic	Item	Checklist item	Section
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable	Subheading 4.1
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up	Subheading 4.1, *see* [87]
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centers	N/A
	5b	Describe eligibility criteria for participants	Subheading 4.1, *see* [87]
	5c	Give details of treatments received, if relevant	N/A
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed	Subheading 4.1
	6b	Report any actions to blind assessment of the outcome to be predicted	N/A

(continued)

Section/topic	Item	Checklist item	Section
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured	2999 predictors, for more details *see* [87]
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors	N/A
Sample size	8	Explain how the study size was arrived at	Subheading 4.1, *see* [87]
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method	The data was already preprocessed and imputed, *see* Subheading 4.1
Statistical analysis methods	10c	For validation, describe how the predictions were calculated	Subheading 3.3
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models	Subheading 3.3
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done	N/A

(continued)

Section/topic	Item	Checklist item	Section
Risk groups	11	Provide details on how risk groups were created, if done	N/A
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors	N/A
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful	Subheading 4.1, *see* [87]
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome	Subheading 4.1, *see* [87]
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors, and outcome)	Subheadings 4.3 and 4.4

(continued)

Section/topic	Item	Checklist item	Section
Model performance	16	Report performance measures (with CIs) for the prediction model	N/A
Model-updating	17	If done, report the results from any model updating (i.e., model specification, model performance)	Subheading 4.4
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data)	Subheading 4.1, *see* [87]
Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data	N/A
	19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence	Subheadings 4.4 and 5
Implications	20	Discuss the potential clinical use of the model and implications for future research	Subheadings 4.4 and 5. However, performance of the model is data-driven
Other information			

(continued)

Section/topic	Item	Checklist item	Section
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, web calculator, and data sets	Subheading 4.1, *see* [87]
Funding	22	Give the source of funding and the role of the funders for the present study	NIH

Selected Open Source (R/Bioconductor/Web-Based) Tools for Supervised Learning Algorithms

Method	Source	Reference
PLS-DA	Bioconductor (ropls)	[93]
PLS-DA, RF, and SVM	Bioconductor (biosigner)	[94]
SVM, RF	Bioconductor (MLSeq)	[95]
RF, SVM, PLS-DA	Metaboanalyst http://www.metaboanalyst.ca/	[96]
PCA, PLS-DA, RF	Bioconductor (statTarget)	[97]
Feature selection, metric evaluation	Bioconductor (OmicsMarker)	[98]
Sparse PLS-DA	Bioconductor (mixOmics)	[99]
Feature selection, metric evaluation	CRAN (lilikoi)	[100]
Probabilistic principal component analysis	CRAN (MetabolAnalyze)	[101]
Kernel-based metabolite differential analysis	CRAN (KMDA)	[21]
PLS-DA, OPLS-DA	CRAN (muma)	[102]
RF	CRAN (RFmarkerDetector)	[103]
RF, SVM, PLS-DA	CRAN (caret)	[104]

References

1. Maniscalco M, Fuschillo S, Paris D, Cutignano A, Sanduzzi A, Motta A (2019) Clinical metabolomics of exhaled breath condensate in chronic respiratory diseases. Adv Clin Chem 88:121–149. https://doi.org/10.1016/bs.acc.2018.10.002
2. Pujos-Guillot E, Petera M, Jacquemin J, Centeno D, Lyan B, Montoliu I, Madej D, Pietruszka B, Fabbri C, Santoro A, Brzozowska A, Franceschi C, Comte B (2018) Identification of pre-frailty sub-phenotypes in elderly using metabolomics. Front Physiol 9:1903. https://doi.org/10.3389/fphys.2018.01903
3. Sarode GV, Kim K, Kieffer DA, Shibata NM, Litwin T, Czlonkowska A, Medici V (2019) Metabolomics profiles of patients with Wilson disease reveal a distinct metabolic signature. Metabolomics 15(3):43. https://doi.org/10.1007/s11306-019-1505-6
4. Wang X, Zhang A, Sun H (2013) Power of metabolomics in diagnosis and biomarker discovery of hepatocellular carcinoma. Hepatology 57(5):2072–2077
5. Caesar LK, Kellogg JJ, Kvalheim OM, Cech NB (2019) Opportunities and limitations for untargeted mass spectrometry metabolomics to identify biologically active constituents in complex natural product mixtures. J Nat Prod 82:469. https://doi.org/10.1021/acs.jnatprod.9b00176
6. Liu LL, Lin Y, Chen W, Tong ML, Luo X, Lin LR, Zhang HL, Yan JH, Niu JJ, Yang TC (2019) Metabolite profiles of the cerebrospinal fluid in neurosyphilis patients determined by untargeted metabolomics analysis. Front Neurosci 13:150. https://doi.org/10.3389/fnins.2019.00150
7. Sanchez-Arcos C, Kai M, Svatos A, Gershenzon J, Kunert G (2019) Untargeted metabolomics approach reveals differences in host plant chemistry before and after infestation with different pea aphid host races. Front Plant Sci 10:188. https://doi.org/10.3389/fpls.2019.00188
8. Wang R, Yin Y, Zhu ZJ (2019) Advancing untargeted metabolomics using data-independent acquisition mass spectrometry technology. Anal Bioanal Chem 411:4349. https://doi.org/10.1007/s00216-019-01709-1
9. Allwood JW, Xu Y, Martinez-Martin P, Palau R, Cowan A, Goodacre R, Marshall A, Stewart D, Howarth C (2019) Rapid UHPLC-MS metabolite profiling and phenotypic assays reveal genotypic impacts of nitrogen supplementation in oats. Metabolomics 15(3):42. https://doi.org/10.1007/s11306-019-1501-x
10. Fang J, Zhao H, Zhang Y, Wong M, He Y, Sun Q, Xu S, Cai Z (2019) Evaluation of gas chromatography-atmospheric pressure chemical ionization tandem mass spectrometry as an alternative to gas chromatography tandem mass spectrometry for the determination of polychlorinated biphenyls and polybrominated diphenyl ethers. Chemosphere 225:288–294. https://doi.org/10.1016/j.chemosphere.2019.03.011
11. Lohr KE, Camp EF, Kuzhiumparambil U, Lutz A, Leggat W, Patterson JT, Suggett DJ (2019) Resolving coral photoacclimation dynamics through coupled photophysiological and metabolomic profiling. J Exp Biol 222:jeb195982. https://doi.org/10.1242/jeb.195982
12. Baumeister TUH, Ueberschaar N, Schmidt-Heck W, Mohr JF, Deicke M, Wichard T, Guthke R, Pohnert G (2018) DeltaMS: a tool to track isotopologues in GC- and LC-MS data. Metabolomics 14(4):41. https://doi.org/10.1007/s11306-018-1336-x
13. Gilmore IS, Heiles S, Pieterse CL (2019) Metabolic imaging at the single-cell scale: recent advances in mass spectrometry imaging. Annu Rev Anal Chem (Palo Alto Calif) 12:201. https://doi.org/10.1146/annurev-anchem-061318-115516
14. Do KT, Wahl S, Raffler J, Molnos S, Laimighofer M, Adamski J, Suhre K, Strauch K, Peters A, Gieger C, Langenberg C, Stewart ID, Theis FJ, Grallert H, Kastenmuller G, Krumsiek J (2018) Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. Metabolomics 14(10):128. https://doi.org/10.1007/s11306-018-1420-2
15. Liggi S, Hinz C, Hall Z, Santoru ML, Poddighe S, Fjeldsted J, Atzori L, Griffin JL (2018) KniMet: a pipeline for the processing of chromatography-mass spectrometry metabolomics data. Metabolomics 14(4):52. https://doi.org/10.1007/s11306-018-1349-5
16. Fielding S, Fayers PM, McDonald A, McPherson G, Campbell MK (2008) Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. Health Qual Life Outcomes 6(1):57

17. Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7(2):147
18. Steyerberg EW, van Veen M (2007) Imputation is beneficial for handling missing data in predictive models. J Clin Epidemiol 60 (9):979
19. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78 (3):779–787. https://doi.org/10.1021/ac051437y
20. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, Ni Y (2018) Missing value imputation approach for mass spectrometry-based metabolomics data. Sci Rep 8(1):663. https://doi.org/10.1038/s41598-017-19120-0
21. Zhan X, Patterson AD, Ghosh D (2015) Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. BMC Bioinformatics 16:77. https://doi.org/10.1186/s12859-015-0506-3
22. Gromski PS, Xu Y, Kotze HL, Correa E, Ellis DI, Armitage EG, Turner ML, Goodacre R (2014) Influence of missing values substitutes on multivariate analysis of metabolomics data. Metabolites 4(2):433–452. https://doi.org/10.3390/metabo4020433
23. Kumar N, Hoque MA, Shahjaman M, Islam SM, Mollah MN (2017) Metabolomic biomarker identification in presence of outliers and missing values. Biomed Res Int 2017:2437608. https://doi.org/10.1155/2017/2437608
24. Sun X, Langer B, Weckwerth W (2015) Challenges of inversely estimating Jacobian from metabolomics data. Front Bioeng Biotechnol 3:188. https://doi.org/10.3389/fbioe.2015.00188
25. Lee JY, Styczynski MP (2018) NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data. Metabolomics 14(12):153. https://doi.org/10.1007/s11306-018-1451-8
26. Di Guida R, Engel J, Allwood JW, Weber RJM, Jones MR, Sommer U, Viant MR, Dunn WB (2016) Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics 12(5):93. https://doi.org/10.1007/s11306-016-1030-9
27. Chen MX, Wang SY, Kuo CH, Tsai IL (2019) Metabolome analysis for investigating host-gut microbiota interactions. J Formos Med Assoc 118(Suppl 1):S10–S22. https://doi.org/10.1016/j.jfma.2018.09.007
28. Shen X, Zhu ZJ (2019) MetFlow: an interactive and integrated workflow for metabolomics data cleaning and differential metabolite discovery. Bioinformatics 35:2870. https://doi.org/10.1093/bioinformatics/bty1066
29. McLachlan, Geoffrey J (2004) Discriminant analysis and statistical pattern recognition. Wiley-Interscience, Hoboken, N.J. John Wiley & Sons. & Wiley InterScience (Online Service)
30. McCallum A, Nigam K (1998) A comparison of event models for naive Bayes text classification. In: AAAI-98 workshop on learning for text categorization, vol 1. Citeseer, pp 41–48
31. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73(16):5261–5267
32. Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106
33. Breiman L (2017) Classification and regression trees. Routledge, Boca Raton
34. Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2 (3):18–22
35. Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recogn Lett 27(4):294–300
36. Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C, Jia W, Zhao A (2013) Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evid Based Complement Alternat Med 2013:298183
37. Scott I, Lin W, Liakata M, Wood J, Vermeer CP, Allaway D, Ward J, Draper J, Beale M, Corol D (2013) Merits of random forests emerge in evaluation of chemometric classifiers by external validation. Anal Chim Acta 801:22–33
38. Ho TK (1998) Nearest neighbors in random subspaces. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, pp 640–648
39. Biau G (2012) Analysis of a random forests model. J Mach Learn Res 13 (Apr):1063–1095
40. Hapfelmeier A, Hothorn T, Ulm K, Strobl C (2014) A new variable importance measure

for random forests with missing data. Stat Comput 24(1):21–34

41. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht FA (2009) A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 10(1):213
42. Maker AV, Hu V, Kadkol SS, Hong L, Brugge W, Winter J, Yeo CJ, Hackert T, Buchler M, Lawlor RT, Salvia R, Scarpa A, Bassi C, Green S (2019) Cyst fluid biosignature to predict Intraductal papillary mucinous neoplasms of the pancreas with high malignant potential. J Am Coll Surg 228:721. https://doi.org/10.1016/j.jamcollsurg.2019.02.040
43. Tkachev V, Sorokin M, Mescheryakov A, Simonov A, Garazha A, Buzdin A, Muchnik I, Borisov N (2018) FLOating-window projective separator (FloWPS): a data trimming tool for support vector machines (SVM) to improve robustness of the classifier. Front Genet 9:717. https://doi.org/10.3389/fgene.2018.00717
44. Yerukala Sathipati S, Ho SY (2018) Identifying a miRNA signature for predicting the stage of breast cancer. Sci Rep 8(1):16138. https://doi.org/10.1038/s41598-018-34604-3
45. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
46. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory. ACM, pp 144–152
47. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
48. Ripley BD (1994) Flexible non-linear approaches to classification. In: From statistics to neural networks. Springer, Berlin, pp 105–126
49. Contreras-Jodar A, Nayan NH, Hamzaoui S, Caja G, Salama AAK (2019) Heat stress modifies the lactational performances and the urinary metabolomic profile related to gastrointestinal microbiota of dairy goats. PLoS One 14(2):e0202457. https://doi.org/10.1371/journal.pone.0202457
50. Park HG, Jang KS, Park HM, Song WS, Jeong YY, Ahn DH, Kim SM, Yang YH, Kim YG (2019) MALDI-TOF MS-based total serum protein fingerprinting for liver cancer diagnosis. Analyst 144:2231. https://doi.org/10.1039/c8an02241k
51. Quiros-Guerrero L, Albertazzi F, Araya-Valverde E, Romero RM, Villalobos H, Poveda L, Chavarria M, Tamayo-Castillo G (2019) Phenolic variation among Chamaecrista nictitans subspecies and varieties revealed through UPLC-ESI(−)-MS/MS chemical fingerprinting. Metabolomics 15(2):14. https://doi.org/10.1007/s11306-019-1475-8
52. Wang J, Yan D, Zhao A, Hou X, Zheng X, Chen P, Bao Y, Jia W, Hu C, Zhang ZL, Jia W (2019) Discovery of potential biomarkers for osteoporosis using LC-MS/MS metabolomic methods. Osteoporos Int 30:1491. https://doi.org/10.1007/s00198-019-04892-0
53. Grissa D, Petera M, Brandolini M, Napoli A, Comte B, Pujos-Guillot E (2016) Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. Front Mol Biosci 3:30. https://doi.org/10.3389/fmolb.2016.00030
54. Bayci AWL, Baker DA, Somerset AE, Turkoglu O, Hothem Z, Callahan RE, Mandal R, Han B, Bjorndahl T, Wishart D, Bahado-Singh R, Graham SF, Keidan R (2018) Metabolomic identification of diagnostic serum-based biomarkers for advanced stage melanoma. Metabolomics 14(8):105. https://doi.org/10.1007/s11306-018-1398-9
55. Catav SS, Elgin ES, Dag C, Stark JL, Kucukakyuz K (2018) NMR-based metabolomics reveals that plant-derived smoke stimulates root growth via affecting carbohydrate and energy metabolism in maize. Metabolomics 14(11):143. https://doi.org/10.1007/s11306-018-1440-y
56. Guo JG, Guo XM, Wang XR, Tian JZ, Bi HS (2019) Metabolic profile analysis of free amino acids in experimental autoimmune uveoretinitis rat plasma. Int J Ophthalmol 12(1):16–24. https://doi.org/10.18240/ijo.2019.01.03
57. Rodrigues-Neto JC, Correia MV, Souto AL, Ribeiro JAA, Vieira LR, Souza MT Jr, Rodrigues CM, Abdelnur PV (2018) Metabolic fingerprinting analysis of oil palm reveals a set of differentially expressed metabolites in fatal yellowing symptomatic and non-symptomatic plants. Metabolomics 14(10):142. https://doi.org/10.1007/s11306-018-1436-7
58. Wong M, Lodge JK (2012) A metabolomic investigation of the effects of vitamin E supplementation in humans. Nutr Metab (Lond) 9(1):110. https://doi.org/10.1186/1743-7075-9-110

59. Li Y, Chen M, Liu C, Xia Y, Xu B, Hu Y, Chen T, Shen M, Tang W (2018) Metabolic changes associated with papillary thyroid carcinoma: a nuclear magnetic resonance-based metabolomics study. Int J Mol Med 41 (5):3006–3014. https://doi.org/10.3892/ijmm.2018.3494
60. Rezig L, Servadio A, Torregrossa L, Miccoli P, Basolo F, Shintu L, Caldarelli S (2018) Diagnosis of post-surgical fine-needle aspiration biopsies of thyroid lesions with indeterminate cytology using HRMAS NMR-based metabolomics. Metabolomics 14(10):141. https://doi.org/10.1007/s11306-018-1437-6
61. Westerhuis JA, van Velzen EJ, Hoefsloot HC, Smilde AK (2010) Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. Metabolomics 6(1):119–128
62. Liquet B, Le Cao KA, Hocini H, Thiebaut R (2012) A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. BMC Bioinformatics 13:325. https://doi.org/10.1186/1471-2105-13-325
63. Liu H, Motoda H (1998) Feature extraction, construction and selection: a data mining perspective, vol 453. Springer Science & Business Media, Norwell
64. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
65. Weston J, Elisseeff A, Schölkopf B, Tipping M (2003) Use of the zero-norm with linear models and kernel methods. J Mach Learn Res 3(Mar):1439–1461
66. Mladenic D, Grobelnik M (1999) Feature selection for unbalanced class distribution and naive bayes. In: ICML 1999, pp 258–267
67. Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52(3):345–370
68. Guan W, Zhou M, Hampton CY, Benigno BB, Walker LD, Gray A, McDonald JF, Fernández FM (2009) Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC Bioinformatics 10(1):259
69. Platt J (1998) Sequential minimal optimization: a fast algorithm for training support vector machines
70. Kuhn M, Johnson K (2013) Applied predictive modeling, vol 26. Springer, New York
71. Behnamian A, Millard K, Banks SN, White L, Richardson M, Pasher J (2017) A systematic approach for variable selection with random forests: achieving stable variable importance values. IEEE Geosci Remote Sens Lett 14 (11):1988–1992
72. Van Calster B, Vickers AJ (2015) Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making 35 (2):162–169
73. Agresti A (2002) Categorical data analysis. Wiley, New York
74. Huang Y, Sullivan Pepe M, Feng Z (2007) Evaluating the predictiveness of a continuous marker. Biometrics 63(4):1181–1188
75. Holder LB, Haque MM, Skinner MK (2017) Machine learning for epigenetics and future medical applications. Epigenetics 12 (7):505–514. https://doi.org/10.1080/15592294.2017.1329068
76. Chen C, Liaw A, Breiman L (2004) Using random forest to learn imbalanced data, vol 110. University of California, Berkeley, pp 1–12
77. Breiman L, Friedman J, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall, New York
78. Japkowicz N (2000) Learning from imbalanced data sets: a comparison of various strategies. In: AAAI workshop on learning from imbalanced data sets. Menlo Park, CA, pp 10–15
79. Maloof MA (2003) Learning when data sets are imbalanced and when costs are unequal and unknown. In: ICML-2003 workshop on learning from imbalanced data sets II, pp 2–1
80. Ling CX, Li C (1998) Data mining for direct marketing: problems and solutions. In: KDD 1998, pp 73–79
81. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
82. Kubat M, Matwin S (1997) Addressing the curse of imbalanced training sets: one-sided selection. In: ICML 1997. Citeseer, pp 179–186
83. Domingos P (1999) Metacost: a general method for making classifiers cost-sensitive. In: KDD 1999, pp 155–164
84. Cateni S, Colla V, Vannucci M (2014) A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing 135:32–41
85. Drummond C, Holte RC (2003) C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Workshop on learning from imbalanced datasets II. Citeseer, pp 1–8

86. Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 13(1):1
87. Cruickshank-Quinn CI, Jacobson S, Hughes G, Powell RL, Petrache I, Kechris K, Bowler R, Reisdorph N (2018) Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. Sci Rep 8(1):17132
88. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD (2010) Genetic epidemiology of COPD (COPDGene) study design. COPD 7(1):32–43. https://doi.org/10.3109/15412550903499522
89. Andersen SL, Briggs FBS, Winnike JH, Natanzon Y, Maichle S, Knagge KJ, Newby LK, Gregory SG (2019) Metabolome-based signature of disease pathology in MS. Mult Scler Relat Disord 31:12–21. https://doi.org/10.1016/j.msard.2019.03.006
90. Lee HS, Seo C, Hwang YH, Shin TH, Park HJ, Kim Y, Ji M, Min J, Choi S, Kim H, Park AK, Yee ST, Lee G, Paik MJ (2019) Metabolomic approaches to polyamines including acetylated derivatives in lung tissue of mice with asthma. Metabolomics 15(1):8. https://doi.org/10.1007/s11306-018-1470-5
91. Long NP, Yoon SJ, Anh NH, Nghi TD, Lim DK, Hong YJ, Hong SS, Kwon SW (2018) A systematic review on metabolomics-based diagnostic biomarker discovery and validation in pancreatic cancer. Metabolomics 14(8):109. https://doi.org/10.1007/s11306-018-1404-2
92. Regan EA, Hersh CP, Castaldi PJ, DeMeo DL, Silverman EK, Crapo JD, Bowler RP (2019) Omics and the search for blood biomarkers in COPD: insights from COPDGene. Am J Respir Cell Mol Biol 61:143. https://doi.org/10.1165/rcmb.2018-0245PS
93. Thévenot EA (2016) ropls: PCA, PLS (-DA) and OPLS (-DA) for multivariate analysis and feature selection of omics data
94. Rinaudo P, Boudah S, Junot C, Thévenot EA (2016) Biosigner: a new method for the discovery of significant molecular signatures from omics data. Front Mol Biosci 3:26
95. Zararsiz G, Goksuluk D, Korkmaz S, Eldem V, Duru IP, Unver T, Ozturk A, Zararsiz MG, klaR M, biocViews Sequencing, R (2014) Package 'MLSeq'
96. Xia J, Psychogios N, Young N, Wishart DS (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res 37(suppl_2):W652–W660
97. Luan H, Ji F, Chen Y, Cai Z (2018) statTarget: a streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data. Anal Chim Acta 1036:66–72
98. Determan Jr CE, Determan Jr MCE (2015) Package 'OmicsMarkeR'
99. Rohart F, Gautier B, Singh A, Le Cao K-A (2017) mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 13(11):e1005752
100. Al-Akwaa FM, Yunits B, Huang S, Alhajaji H, Garmire LX (2018) Lilikoi: an R package for personalized pathway-based classification modeling using metabolomics data. GigaScience 7(12):giy136
101. Gift N, Gormley IC, Brennan L, Gormley MC (2010) Package 'MetabolAnalyze'
102. Gaude E, Chignola F, Spiliotopoulos D, Spitaleri A, Ghitti M, Garcìa-Manteiga JM, Mari S, Musco G (2013) Muma, an R package for metabolomics univariate and multivariate statistical analysis. Curr Metabol 1(2):180–189
103. Palla P (2015) Information management and multivariate analysis techniques for metabolomics data. Universita'degli Studi di Cagliari
104. Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28(5):1–26