npj Precision Oncology

https://doi.org/10.1038/s41698-026-01342-z

Article in Press

# Ensemble learning on serum metabolic fingerprints for early detection of lung adenocarcinoma

Received: 26 June 2025

Accepted: 17 February 2026

Cite this article as: Cai, C., Xu, W., Yang, S. *et al.* Ensemble learning on serum metabolic fingerprints for early detection of lung adenocarcinoma. *npj Precis. Onc.* (2026). https://doi.org/10.1038/s41698-026-01342-z

Chenlei Cai, Weijie Xu, Shuo Yang, Jia Yu, Lei Wang & Shengxiang Ren

# Ensemble learning on serum metabolic fingerprints for early detection of lung adenocarcinoma

Chenlei Cai[1,3*], Weijie Xu[2,3], Shuo Yang[1], Jia Yu[1], Lei Wang[1*], and Shengxiang Ren[1*]

1 Department of Medical Oncology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai 200433, China.

2 Department of Clinical Laboratory, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai 200433, China

3 Chenlei Cai and Weijie Xu contributed equally to this work.

Corresponding authors:

Email: harry_ren@tongji.edu.cn (Shengxiang Ren); wangleixxxn@163.com (Lei Wang); chenlei_cai@tongji.edu.cn (Chenlei Cai)

**Abstract**

Lung adenocarcinoma (LUAD) remains a leading cause of cancer-related mortality worldwide, highlighting the urgent need for non-invasive strategies for early detection. Here, we present a machine learning-assisted metabolomics approach for the early detection of LUAD. Untargeted metabolomic profiling was performed on 199 serum samples from healthy individuals, patients with lung precancerous lesions, and those with stage I LUAD. An ensemble machine learning workflow was developed to identify metabolite panels capable of discriminating clinical status with high accuracy. We observed progressive metabolic alterations in bile acid, lipid, amino acid, and purine metabolism during LUAD initiation and stepwise progression. Notably, ensemble learning identified a six-metabolite panel, including 12-hydroxydodecanoic acid, hypoxanthine, xanthosine, cholic acid, agmatine, and paraxanthine, for accurate detection of early-stage LUAD, and a distinct four-metabolite panel, comprising 7-α,27-dihydroxycholesterol, 11-undecanedicarboxylic acid, biliverdin, and Prolyl-Valine, for precise differentiation between pre-invasive and invasive lesions. Both panels demonstrated promising diagnostic potential, with performance metrices comparing favorably to established methodologies within the current study cohort. This study delineates the evolutionary trajectory of the serum metabolome associated with early LUAD pathogenesis and provide promising biomarkers for non-invasive early detection.

**Keywords** Lung Adenocarcinoma, Metabolomics, Early Diagnosis, Serum, Machine Learning

## INTRODUCTION

Lung cancer remains the most commonly diagnosed cancer and the leading cause of cancer death globally [1]. Among them, lung adenocarcinoma (LUAD) is the most common subtype [2]. Although low-dose computed tomography (LDCT) screening is recommended for high-risk populations, the management of pulmonary nodules detected by LDCT, which are the primary indication of early-stage LUAD, is challenging due to a high false-positive rate up to 96% [3]. Furthermore, it is difficult for LDCT to assess the malignant risk of indeterminate pulmonary nodules, which account for around 50–76% [4]. Therefore, there is an urgent, unmet need for the development of novel non-invasive methods for early detection of LUAD.

Comprehensive measurement of circulating molecules in blood plasma and/or serum, including genomics, epigenomics, and proteomics, has gained increasing attention in discovering potential biomarkers for cancer diagnosis [5–8]. Meanwhile, metabolomics can provide insights into the cellular processes in response to the influence of genetic and environmental risk factors. Serum metabolomics has attracted creasing interest as a minimally invasive and sensitive platform, with metabolites being closer to the phenotype and more abundant than circulating DNA or protein biomarkers, enabling early detection of LUAD and ensemble learning-based biomarker discovery. Therefore, it has been extensively applied to unravel molecular mechanism, and identify molecular indicators for early detection and prognosis of various diseases [9–15]. For example, several circulating serum/plasma metabolites were found to be helpful in the discrimination between lung cancer and healthy individuals [14–25]. However, most of them were screened by a limited class of known metabolites, lipids, or amino acid profiles [21–23,25]. Meanwhile, the clinical decision-

making in different adenocarcinoma classifications further strengthen the necessary to distinguish the early LUAD from its precursors [21–25]. Therefore, the global metabolomic profiling of plasma/serum with high coverage is highly desired for screening and validating metabolic biomarkers. Recently, artificial intelligence, has been widely used to automatically analyze complex data generated in studies ranging across the biological and biomedical sciences [26–33]. Unlike traditional machine learning approaches that try to learning a single hypothesis from train dataset, ensemble learning algorithms develop multiple hypotheses and combine them to solve a specific issue [34–38].

Here we performed untargeted metabolomics profiling on 199 serum samples including healthy controls (HC), LUAD and its precursors to identify the metabolic changes (**Fig. 1**). We observed evolutionary dynamics in the serum metabolome profiles from HC to pre-invasive adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and subsequent invasive adenocarcinoma (IAC). Moreover, we leveraged ensemble learning to analyze the serum metabolic fingerprints, and uncovered different panels of circulating serum metabolites.

## RESULTS

### Metabolic profiles of early LUAD and preneoplasia

To clarify the serum metabolome profiles during stepwise progression of early-stage LUAD from pre-invasive status to invasive adenocarcinoma (IAC), we conducted high-resolution, non-targeted liquid chromatography-mass spectrometry (LC-MS)-based metabolomics analysis of blood serum from a total of 199 participants including 50 IAC, 49 MIA, 50 AIS, and 50 HC. The

clinical characteristics of the participants are shown in **Fig. 2a** and **Table S1**. In total, 993 metabolites, mainly including lipids, organic acids, amines, amino acids, carbohydrates, nucleosides, purines, and about 28% metabolites of other classes, were detected in the serum samples (**Fig. 2b**). Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) clearly distinguished minimally invasive and invasive LUAD from pre-invasive AIS and healthy controls, and this separation was largely consistent across different data preprocessing methods (**Figs. 2c and S1**). In addition, the high reproducibility of pooled quality control (QC) samples confirmed system stability and reliability of the metabolomics measurements (**Figs S2–S5**). Intriguingly, we found 290 significantly differential metabolites in IAC versus HC (**Fig. 2d** and **Table S2**), whereas 100 differential metabolites in IAC versus AIS, and 11 differential metabolites in IAC versus MIA (two-sided t-test with Welch's correction, FDR < 0.05 and fold change > 1.25 or < 0.8), implying gradual changes in metabolic processes during the stepwise progression of early LUAD (**Figs. 2d** and **S6**). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway-based analysis by MetaboAnalyst revealed several metabolic pathways such as primary bile acid biosynthesis, steroid hormone biosynthesis, purine metabolism, steroid hormone biosynthesis, tryptophan metabolism, and fatty acid biosynthesis, were progressively disturbed in the preneoplastic phase and the early invasive LUAD, compared to the healthy individuals. Additionally, the porphyrin metabolism, glycerophospholipid metabolism, and cysteine and methionine metabolism became perturbed early in the pre- and minimally invasive phases (**Fig. 2e and S7**). Our analysis of circulating serum metabolites is consistent with previously reported results of single-cell and tissue metabolomics studies,[14,22] highlighting that the molecular and metabolic alterations begin earlier

during LUAD initiation and progression.

To characterize the dynamic changes in terms of the serum metabolic landscape from HC to AIS, MIA, and then IAC, we examined the metabolic alterations among the four clinical statuses using the one-way analysis of variance (one-way ANOVA) method (FDR < 0.05), and acquired 146 differential metabolites. We next evaluated the relative abundance distribution of these metabolites with the Euclidean-Ward clustering method (**Fig. 3a**). The metabolites were categorized into two major groups, of which the molecules in the first group showed an overall increase in abundance with the disease progression, while the molecules in the second group showed an overall trend of decreasing concentration. Further analysis of the deregulated metabolites using fuzzy c-means clustering (Mfuzz) revealed six distinct temporal trends (Trends 1-6) associated with LUAD onset and progression (**Figs. 3b, S8 and S9**). Metabolites in Trends 1 and 5 showed a progressive downregulation from healthy controls (HC) to invasive adenocarcinoma (IAC), whereas those in Trend 3 displayed a gradual upregulation. Metabolites in Trend 2 were predominantly elevated in AIS, while Trend 4 metabolites exhibited a marked increase in MIA. Interestingly, metabolites in Trends 6 demonstrated a fluctuating pattern, showing an initial decrease in IAC, an increase in MIA, followed by another decline in IAC. Notably, several metabolites in Trend3 and Trend 5 showed stepwise decreases or increases from HC to through preinvasive to invasive stages, reflecting continuous metabolic reprogramming during early carcinogenesis. The relative levels of PC(18:0/18:1(9Z)), cholesterol sulfate, and Cer(d18:0/14:0) in fatty acid metabolism were gradually increased, whereas dodecanoic acid involved in biosynthesis of unsaturated fatty acids

was decreased. We also observed a moderate increase of biliverdin in porphyrin metabolism, and a steady decrease of serum phenylalanylserine involved in phenylalanine metabolism. In addition, which are involved in purine metabolism and tryptophan metabolism, respectively, also showed a steady downward trend (**Fig. 3c and S9**). Together, these results revealed metabolic perturbations and unique metabolic vulnerabilities at different stages of early LUAD which could be targeted for disease monitoring and detection.

**Metabolite Panels for Patient Diagnosis at Different Stages**

We then leveraged the obtained reprogrammed metabolites to develop innovative non-invasive diagnostic methods for early LUAD at different stages. The data at each stage were randomly split into a training set and test set (80% and 20%, respectively). LASSO regression analysis was first conducted on the training set to select essential metabolites between two given groups (HC versus IAC, HC versus AIS/MIA/IAC, AIS versus MIA/IAC, and MIA versus IAC). The automatic ensemble machine learning algorithm AutoGluon was then used to build a model for predicting the clinical status of each patient. A six-metabolite panel enabling the accurate discrimination between healthy individuals and IAC patients was identified (**Fig. 4a–c**). Specifically, the binary classification of HC versus IAC using the model resulted in a balanced accuracy of 0.95, MCC of 0.905, AUROC of 0.97, F1 of 0.947, precision of 1.00, and recall of 0.90 on the on the test set (**Figs. 4d, S10 and S11**). Consistent model performance was also observed under standard 5-fold cross-validation (AUROC = 0.97, accuracy = 0.95 on the test set), supporting the effectiveness of AutoGluon's built-in ensemble strategy in prevent overfitting (**Fig.**

**S12**). Nate that, pairwise DeLong's and permutation tests did not show statistically significant differences ($p > 0.05$), most likely because of the limited sample size of the independent test cohort, which reduces the statistical power to detect small performance differences. SHapley Additive exPlanations (SHAP) values were used to evaluate the influence of individual metabolites on the predictive performance of the model. Deoxycorticosterone and PC(18:0/18:1(9z)) were found to be the top two important features contributing to the prediction of IAC, while others had a relatively even contribution to the output of the eight-metabolite diagnostic model (**Fig. 4e**). Previous studies on LUAD have consistently identified remarkably perturbed lipids [20,22,39–41]. Significant alterations in deoxycorticosterone and PC(18:0/18:1(9z)) has been detected in blood serum and plasma samples of LUAD patients [40,41]. 25-Hydroxycholesterol has been found to promote migration and invasion of lung adenocarcinoma cells [42]. Decreased serum levels of 2-hydroxyestradiol has been observed in non-small cell lung cancer [43]. Likewise, the relative abundance plots indicated that all of the six metabolites were significantly different between HC and IAC, with two of them (PC(18:0/18:1(9z)) and PC(40:6)) being significantly elevated in IAC and the other four compounds (deoxycorticosterone, 25-Hydroxycholesterol, 2-hydroxyestradiol, and LysoPC(O-18:0/0:0)) being significantly downregulated in IAC (**Fig. 4c**).

To intuitively demonstrate the predictive performance of the six-metabolite model, we generated a scatter plot to compare predicted score of each participant along with his/her actual disease status (IAC/HC). Using a cutoff-value of 0.5 for classification, the model accurately identified all of the IAC patients on the test set (**Fig. 4f**). The diagnostic performance of the six-

metabolite model was comparable to that of other existing metabolic diagnostic strategies for early invasive LUAD (AUC: 0.97 versus 0.894–0.93, **Table S3**). The reasons for our model's promising predictive ability were attributed to the following aspects: first, we performed high-resolution UPLC-MS-based high-coverage global metabolomics measurements to acquire system-wide characterizations of metabolic status in the serum samples of patients and controls; second, we conducted a rigorous multi-step feature selection pipeline—t test followed by LASSO regression analysis to identify the potential metabolic biomarkers. Furthermore, one strength of the AutoGluon library used in our study is its ensemble learning ability, which combines predictions from multiple models to improve accuracy and robustness. **Fig. S10** shows the IAC prediction accuracies of different models trained by the AutoGluon Tabular Prediction function. As expected, the weighted ensemble model performed best on both the internal validation set and the external test set. As a control experiment, we applied the AutoGluon Tabular Prediction function to retrain classification models after removing the multilayer stacking parameter in the script. In this case, the accuracy of the best ensemble model decreased to 0.90 on the test set (**Fig. S13**). Furthermore, we also benchmarked the performance of the established metabolic diagnostic model with different machine learning algorithms in MetaboAnalyst, including supporting vector machine, PLS-DA, random forest, and logistic regression. The ensemble learning model showed the best performance for IAC identification (**Fig. S14**). Together, the above data indicated that our assembled model based on a panel of eight differential metabolites could achieve accurate diagnosis of early-stage IAC.

Moreover, for the discrimination of HC from AIS/MIA/IAC, six metabolites including 12-hydroxydodecanoic acid, hypoxanthine, xanthosine, cholic acid, agmatine, and paraxanthine generated an excellent classification accuracy of 1.00 on the test data (**Fig. 5a-d**). An increase in cholic acid, and a decrease in other five metabolites were found in the disease group (**Fig. 5c**). SHAP analysis revealed that the six differential metabolites contributed relatively equally to the predictive performance of the model (**Fig. 5e**). Further observation of the distribution of samples in each group revealed that the samples in the HC group were highly clustered, while those in the LUAD group were relatively divergent, which may be due to the fact that the LUAD group was composed of individuals with different disease states, resulting in more obvious intra-group heterogeneity (**Fig. 5f**).

In addition, a four-metabolite panel consisting of 7α,27-dihydroxycholesterol, 1,11-undecanedicarboxylic acid, biliverdin, and prolyl-valine were found to be effective in distinguishing between AIS and MIA/IAC with a high AUROC value of 1.00 on the test set. Among the four metabolites, 7a,27-dihydroxycholesterol, 1,11-undecanedicarboxylic acid, and prolyl-valine were decreased whereas biliverdin was increased in the invasive group ((**Fig. 6**). The differentiation between MIA and AIS is beneficial to reduce over diagnosis. MIA requires surgical intervention due to its invasiveness and worse prognosis, while AIS as pre-invasive nodules only need regular follow-ups. Currently, it is challenging to accurately distinguish AIS from MIA by chest computed tomography scans in clinical practice [44]. The above results demonstrated the promise of applying machine learning analysis of metabolomics data, which facilitated the

discovery of circulating markers for non-invasive early detection of LUAD.

Although both MIA and IAC are early LUAD, their invasiveness and prognosis are quite different. With a 5-year disease-free survival rate of nearly 100%, MIA had a much better prognosis compared with that of stage I IAC, which was down to about 74.6% [45]. Different surgical treatments are clinically adopted for these two invasive nodules. IAC should be resected with a lobectomy accompanied by the removal of the surrounding lymph nodes. Meanwhile, limited/sublobar resection with optional lymph node resection has been suggested for MIA [46]. Therefore, the preoperative classification of pulmonary nodule pathological subtypes is considered important but challenging for determining the precise surgical procedure. By executing the above metabolic feature selection and analysis pipeline, we identified a panel consisting of two metabolites to enable discrimination of MIA and IAC cases. The binary classification model provided AUROC values of 0.74 and 0.78 on the training and independent test sets (**Fig. 7a**). Statistical analysis found that the serum levels of beta-glycerophosphoric acid and cytidine were significantly decreased in the IAC group compared with that in the MIA group (**Fig. 7b**).



## DISCUSSION

This study integrates global serum metabolomic profiling and advanced machine learning strategies to characterize the temporal metabolic reprogramming associated with the initiation and progression of early-stage LUAD. By leveraging a robust ensemble machine learning framework, we systematically analyzed high-dimensional metabolomic data and identified a rigorously validated six-metabolite panel (12-hydroxydodecanoic acid, hypoxanthine, xanthosine, cholic acid,

agmatine, and paraxanthine) that shows competitive advantages relative to most existing metabolomics-based diagnostic models in detecting early-stage LUAD (**Table S3**). These metabolites reflect key alterations in lipid, nucleotide, and amino acid metabolism, underscoring the extensive biochemical remodeling that accompanies early tumorigenesis. Specifically, elevated hypoxanthine and xanthosine levels suggest enhanced purine turnover that is established as a hallmark of proliferative metabolic demands[47]. Cholic acid and 12-hydroxydodecanoic acid are associated with disruptions in bile acid and fatty acid metabolism, patterns that have been consistently observed in the metabolic rewiring of multiple cancer types, including lung cancer[48-50]. In addition, agmatine has been implicated in cell proliferation and stress response [51], while paraxanthine may indicate altered caffeine metabolism and oxidative status. Our observation aligns with recent findings by Dong et al., who used Mendelian randomization to reveal that dysregulated plasma paraxanthine is associated with lung cancer pathogenesis[52]. Although some metabolites included in the diagnostic panels (e.g., hypoxanthine and xanthosine) have been previously associated with LUAD, our study provides a data-driven and integrative perspective by constructing optimized, interpretable biomarker panels that capture stage-specific metabolic alterations across the LUAD progression spectrum. Compared with previous studies, our work integrates ensemble learning with interpretable feature selection, systematically evaluates robustness across preprocessing methods, and profiles stage-specific metabolic alterations, thereby providing methodological advancement, novel biological insights, and clinically relevant, stable predictive panels.

As the most prevalent histological subtype of lung cancer, LUAD contributes substantially to global cancer-related mortality. This underscores an urgent need to characterize longitudinal metabolic alterations associated with disease progression and develop blood-based biomarkers that facilitate earlier detection. While previous studies have predominantly focused on genomic, transcriptomic, and proteomic changes [53,54], our study address a critical gap by highlighting metabolic remodeling as an early and actionable event in LUAD tumorigenesis. Using LC-MS-based metabolomics, a platform renowned for its analytical robustness, we mapped global metabolite changes in sera across the spectrum of LUAD development, from healthy individuals to those with precancerous lesions and stage I disease. Our findings not only reaffirm the involvement of bile acid, lipid, amino acid, and purine metabolism in LUAD pathogenesis, but also uncover novel metabolite markers capable of distinguishing between pre-invasive and invasive stages. Accurate discrimination of progression stages, such as AIS, MIA and IAC, is clinically critical, as it directly informs treatment decisions, surveillance strategies, and prognostic evaluation[55]. However, current non-invasive modalities such as imaging often fall short in resolving these lesions with sufficient specificity and sensitivity, leading to overtreatment or delayed intervention[56]. Although several efforts have explored molecular features associated with LUAD progression, most focus on tissue-based omics or require invasive sampling[57,58]. Serum metabolomics offers a promising alternative, yet few studies to date have systematically addressed its potential for fine-stage resolution. Our results provide periluminally evidence that serum metabolic profiling may fill this gap, providing a minimally invasive approach to stratify patients by pathological stage.

Despite the advantages of metabolomics in capturing systemic biochemical changes, the high dimensionality and complexity of the data present challenge for clinical deployment. In this context, machine learning methods have emerged as powerful tools to model nonlinear relationships, reduce dimensionality, and improve the predictive power of diagnostic signatures. In our study, machine learning was used across three pivotal stages: First, LASSO regression was employed to minimize overfitting and isolate the most informative features, forming the foundation of our six- and four-metabolite models. Second, we adopted the AutoGluon framework, a robust ensemble learning platform, to further optimize model generalizability and minimize variance during classifier development. Third, this integrative approach enabled the discovery of informative metabolites that might have been overlooked using conventional statistical methods, thereby broadening the repertoire of candidate biomarkers.



Notably, the ensemble learning model identified subtle yet reproducible metabolite alterations between healthy controls and early LUAD, highlighting its sensitivity in capturing early disease signals. The biomarker panels were developed for distinct pairwise comparisons representing different stages of LUAD progression. Among them, the HC vs (AIS + MIA + IAC) model serves as the primary diagnostic classifier, while the other panels provide stage-specific insights into metabolic evolution from preinvasive to invasive adenocarcinoma. The AutoGluon-based approach, which aggregates multiple base learners through weighted ensembling, demonstrated competitive diagnostic capacity relative to conventional algorithms such as logistic regression and random forest, while showing consistent robustness in feature selection. This methodological

framework may serve not only as foundation for future LUAD diagnostic strategies but also a generalizable model for biomarker discovery in other malignancies.

Nevertheless, several limitations warrant consideration. First, the single-center design and relatively small cohort scale may yield optimistic performance metrics which should be interpreted with caution. Consequently, future validation across larger, multi-center cohorts is essential to confirm the generalizability and clinical robustness of these panels. Second, potential confounding effects from demographic variables such as ethnicity, age, and comorbidities were not extensively evaluated and should be addressed in future studies. Third, our results suggest that single-omics serum metabolomics may have limited discriminative power in distinguishing closely related lesion types, such as MIA and IAC. To address this, future research should explore integrative multi-omics strategies combining transcriptomic, proteomic, and epigenomic data to construct combinatorial biomarker panels with enhanced diagnostic resolution.



In summary, this study elucidates the metabolic landscape underlying early LUAD development and present ensemble machine learning-based diagnostic models with high sensitivity and specificity. These findings contribute to a deeper understanding LUAD pathological processes and lay the groundwork for the development of minimally invasive tools for early detection and clinical decision support.

## Methods

### Serum sample collection

Peripheral venous blood-derived serum samples were collected from a total of 199 participants in Shanghai Pulmonary Hospital, which contained 50 HC, 50 patients with AIS, 49 patients with MIA, and 50 patients with IAC. The serum collection criteria for all subjects were as follows: fast for at least 6 h before blood draw. The blood was then clotted for 30 min at room temperature, followed by centrifugation at 3000 rpm for 10 min at 4℃. The supernatant was further centrifuged at 12000 rpm for 10 min at 4℃. The prepared serum was aliquoted and immediately stored at -80°C until metabolic analysis. Informed consents were obtained from each participant before the initiation of the study. The study was conducted in accordance with the principles of the Declaration of Helsinki and approved and approved by the Institute of the Shanghai Pulmonary Hospital of Tongji University (ethics approval number: L20-337-1). Informed consent was obtained from all participants in the study.

**Metabolite extraction**

All serum samples were thawed slowly on ice. For non-targeted analysis, 100 μL serum for each subject was added to 400 μL of pre-cooled methanol/acetonitrile (1:1, v/v). The mixture was kept for 1 min with vortex, and 30 min with sonication at 4℃. Then, the sample was left at -20℃ for 1 h to precipitate proteins. Afterwards, the sample mixture was centrifuged for 20 min at 4℃ with a speed of 14000 rcf to collect the supernatant which was subjected to freeze-drying and immediately stored at -80℃ until analysis. Finally, the sample was dissolved in 500 μL of methanol/$H_2O_2$ (1:1, v/v) for metabolic analysis. For quality control (QC) samples, 10 μL serum of each subject was mixed and then processed the same as that of the study serum samples.

**Untargeted metabolomics**

For untargeted metabolomics, samples were analyzed using an AB SCIEX Triple TOF 6600 mass spectrometer coupled to an Agilent 1290 Infinity ultra-performance liquid chromatography (UPLC) employing a ACQUITY UPLC BEH Amide column (2.1 × 100 mm, 1.7 μm, Waters), and the analysis was conducted by Shanghai Applied Protein Technology (APTBIO). The column temperature and autosampler temperature was set to 25℃ and 45℃. The flow rate was set to 0.5 mL/min, and the injection volume was 2 μL. Mobile phase A was water (with 25 mM ammonium acetate and 25 mM ammonium hydroxide), and mobile phase B was acetonitrile. The gradient was set as follows: 0.0-0.5 min, 95%B; 0.5-7.0 min, linear decrease to 65%B; 7.0-8.0 min, linear decrease to 40% B; 8.0-9.0 min, 40% B; 9.0-9.1 min, linear gradient back to 95%B; 9.1-12.0 min, 95% B. The samples were randomized to avoid the influence of instrument fluctuation on the result. The QC samples were regularly inserted in the test sample to monitor the precision and stability of the method during operation.



The electrospray ionization (ESI) source parameters were set as follows: source temperature 600℃, spray voltage 5.5 kV (for positive mode) and -5.5 kV (for negative mode), atomized gas (Gas1) 60 psi, heated gas (Gas2) 60 psi, and curtain gas (CUR) 30 psi. In MS only acquisition, the instrument was set to acquire over the mass-to-charge (m/z) range 60-1000 Da, and the accumulation time for the TOF MS scan was set at 0.20s/spectrum. In auto MS/MS acquisition, the instrument was set to acquire the m/z range 25-1000 Da, and the accumulation time for the TOF MS scan was set at 0.05s/spectrum. The product ion scan is acquired using information dependent acquisition (IDA) with high sensitivity mode selected. The parameters were set as

follows: collision energy (CE) 35 ± 15 eV; declustering potential (DP) 60 V (for positive mode) and -60 V (for negative mode); isotopes within 4 Da were excluded; and 10 candidate ions were monitored per cycle.

**Data preprocessing and analysis**

Raw LC-MS data were first converted to mzXML files using ProteoWizard MSConvert and then imported into the open-source XCMS software. For peak detection, the centWave algorithm was applied with the following parameters: mass accuracy m/z = 10 ppm, peakwidth = c(10, 60), and prefilter = c(10, 100). Peaks were grouped using bw = 5, mzwid = 0.025, and minfrac = 0.5. CAMERA (Collection of Algorithms of MEtabolite pRofile Annotation) was used for annotation of isotopes and adducts. Only features with more than 50% nonzero measurements in at least one study group were retained for downstream analysis. For metabolite identification, MS1 accurate mass (<10 ppm) and MS/MS spectra were matched against an in-house reference database maintained by APTBIO. This database contains authentic chemical standards with known retention times, accurate masses, MS/MS fragmentation patterns, and collision energy information. The mirror plots comparing experimental MS2 spectra of metabolites with the corresponding reference spectra have been deposited in figshare: (DOI: 10.6084/m9.figshare.30464846).

All serum metabolomics data from 199 participants were processed in MetaboAnalyst 6.0 (http://www.metaboanalyst.ca) for exploratory statistical analysis, including identification of differentially expressed metabolites and pathway enrichment. For this purpose, MS1 signal intensities were first normalized by the sum of all detected peaks within each sample (sum

normalization, equivalent to total ion current normalization at the data matrix level), followed by log10 transformation and auto scaling. These preprocessing steps are widely used as standard and effective approaches in untargeted metabolomics, allowing minimization of systematic variation, correction for sample dilution effects, and placement of all metabolites on a comparable scale. To evaluate the robustness of normalization and scaling, additional analyses were conducted using group-based and single-QC-based probabilistic quotient normalization (PQN), as well as Pareto, Range, and Mean scaling. Comparable PLS-DA score plots, 5-fold cross-validation (R2, Q2) values, and permutation test results confirmed that the findings were not sensitive to the normalization or scaling approach (Fig. S1). These procedures were used solely for the exploratory characterization of global metabolic alterations. Unless otherwise stated, differential metabolites were analyzed by using two-tailed t-test with Welch's correction (FDR < 0.05, and fold change > 1.25 or < 0.8). Clustering of differential metabolites was conducted with the R package 'Mfuzz' (v2.64.0). Pathway enrichment analysis based on the differential metabolites was performed using MetaboAnalyst 6.0. Metabolites that passed the above t-test and fold-change thresholds were mapped to KEGG compound IDs, and their relative abundance values (normalized peak intensities) were used as input for the Quantitative Enrichment Analysis (QEA) module within MetaboAnalyst. QEA leverages quantitative metabolite measurements to assess pathway enrichment; in this way the KEGG human pathway library was used as the reference. Pathways with FDR < 0.05 (after multiple testing correction within MetaboAnalyst) were considered significantly enriched.

**Machine learning analysis**

The serum metabolite biomarker model for prediction of disease stages was established using the AutoGluon tabular predictor (v1.1.0), with feature selection performed via a combination of Least Absolute Shrinkage and Selection Operator (LASSO) and t-test. For each metabolite panel, the participants were randomly stratified sampling into a discovery dataset (80%) and test dataset (20%). Next, LASSO regression was performed using the R package glmnet (v4.1-7) to select a reduced set of features. These features were then intersected with those identified by t-test (FDR $< 0.05$) to distinguish patients at different clinical status. The alpha parameter was set to 1 and five-fold cross-validation was used with the minimum criteria in the LASSO analysis. The nonzero coefficients were selected to identify differential metabolic features. Then, AutoGluon-based assembling learning was conducted with the selected features in the discovery dataset to train multiple different ML models in parallel, and finally to build a weighted assembled model with the best prediction performance in a Jupyterlab environment (v4.0.8). The basic model algorithms included random forest, CatBoost, XGBoost, LightGBM, ExtraTrees, and so on. Afterward, the diagnostic model was applied to the test dataset. For each ML model, accuracy, area under curve (AUC), precision, sensitivity/recall, specificity, F1 score, and the Matthews correlation coefficient (MCC) were calculated by the tabular predictor. Model robustness was evaluated using both AutoGluon's internal bagging-based validation and repeated 10 × 5 stratified cross-validation. AutoGluon's bagging configuration (num_bag_folds = 8, num_bag_sets = 5) enables each base model to utilize 87.5% of the training data while introducing stochastic resampling to enhance ensemble diversity and bias-variance optimization (See key parameters in the Table S4). The internal validation score reflects out-of-fold performance within the bagging process, whereas

repeated cross-validation provided an independent and more conservative estimate of model stability (mean AUC = 0.918 ± 0.084). The final ensemble model achieved an AUC of 0.97 and accuracy of 0.95 on the test set, confirming good generalization. SHAP values were used to interpret model predictions and identify influential metabolites. Notably, while SHAP indicates statistical importance in the predictive model, biological relevance was inferred based on consistency with univariate results, pathway enrichment, and literature evidence, rather than SHAP ranking alone. Receiver operating characteristic (ROC) and confusion matrix were generated with the scikit-learn package (v1.3.2). Addition Python libraries used to support data analysis and visualization included pandas (v. 2.1.4), numpy (v. 1.21.5), scipy (v1.14.1) SHAP (v0.44.0), and seaborn (v0.12.2).



**Statistical analysis**

Statistical analysis methods used for metabolomics analysis and model evaluation were described in the 'Results/Discussion' section, corresponding figure legends, and 'Methods' subsections. Specifically, two-sided Welch's t-test was used for comparing the means of two samples with unequal variances. Adjusted P-value of less than 0.05 was considered statistically significant. OriginPro (v2021b), R (v4.4.0) software (https://cran.r-project.org/), and Python (v3.10.1) were used to perform data analysis.

**Supplementary information**

The online version contains supplementary material available at https://doi.org/.

**Author contributions**

CC and SR designed the study. Data collection was carried out by CC, WX, and LW. Statistical analysis and graph organization were made by CC and WX. The initial manuscript was written by CC, WX, LW, SY and JY. A manuscript review was made by CC, WX, LW, and SR.

**Acknowledgements**

This study was supported by the National Natural Science Foundation of China (82404096), the Science and Technology Commission of Shanghai Municipality (24Y12800300) and the National Key Clinical Specialty Discipline Construction Program of China: Establishment and Application of a Precision Diagnosis and Treatment System for Chest Tumors.

**Availability of data and materials**

All the data and materials that support the findings of this study are available within the article and supplemental information or available from the authors upon request.

**Competing interests**

The authors declare no competing interests.

**References**

(1) Siegel, R. L. et al. Cancer Statistics. *CA. Cancer. J. Clin.* **72**, 7–33 (2022).

(2) Cheng, T. Y. et al. The International Epidemiology of Lung Cancer: Latest Trends, Disparities, and Tumor Characteristics. *J. Thorac. Oncol.* **11**, 1653–1671 (2016).

(3) Nair, A. et al. Variable Radiological Lung Nodule Evaluation Leads to Divergent Management Recommendations. *Eur. Respir. J.* **52**, 1801359 (2018).

(4) Vachani, A. et al. Factors That Influence Physician Decision Making for Indeterminate Pulmonary Nodules. *Ann. Am. Thorac. Soc.* 11,1586–1591 (2014).

(5) Crowley, E. et al. Liquid Biopsy: Monitoring Cancer-Genetics in the Blood. *Nat. Rev. Clin. Oncol.* **10**, 472-84 (2013).

(6) Fedyuk, V. et al. Multiplexed, Single-Molecule, Epigenetic Analysis of Plasma-Isolated Nucleosomes for Cancer Diagnostics. Nat. Biotechnol. 41, 212–221(2023).

(7) Gardner, L. et al. Nano-Omics: Nanotechnology-Based Multidimensional Harvesting of the Blood-Circulating Cancerome. *Nat. Rev. Clin. Oncol.* **19**, 551–561 (2022).

(8) Hu, A. et al. Cancer Serum Atlas-Supported Precise Pan-Targeted Proteomics Enable Multicancer Detection. Anal. Chem. **95**, 862–871 (2023).

(9) Buergel, T. et al. Metabolomic Profiles Predict Individual Multidisease Outcomes. *Nat. Med.* **28**, 2309–2320 (2022).

(10) Chen, F. et al. Integrated Analysis of the Faecal Metagenome and Serum Metabolome Reveals the Role of Gut Microbiome-Associated Metabolites in the Detection of Colorectal Cancer and Adenoma. *Gut.* **71**, 1315–1325 (2022).

(11) Yi, R. et al. Multi-Omic Profiling of Multi-Biosamples Reveals the Role of Amino Acid and Nucleotide Metabolism in Endometrial Cancer. *Front. Oncol.* **12**, 861142 (2022).

(12) Sinclair, E. et al. Metabolomics of Sebum Reveals Lipid Dysregulation in Parkinson's Disease. *Nat. Commun.* **12**, 1592 (2021).

(13) Wang, Y. et al. Self-Assembled Hyperbranched Gold Nanoarrays Decode Serum United Urine Metabolic Fingerprints for Kidney Tumor Diagnosis. *ACS. Nano.* **18**, 2409–2420 (2024).

(14) Wang, G. et al. Lung Cancer Scrna-Seq and Lipidomics Reveal Aberrant Lipid Metabolism for Early-Stage Diagnosis. *Sci. Transl. Med.* **14**, eabk2756 (2022).

(15) You, L. et al. Liquid Chromatography-Mass Spectrometry-Based Tissue Metabolic Profiling Reveals Major Metabolic Pathway Alterations and Potential Biomarkers of Lung Cancer. *J. Proteome. Res.* **19**, 3750–3760(2020).

(16) Mathé, E. A. et al. Noninvasive Urinary Metabolomic Profiling Identifies Diagnostic and Prognostic Markers in Lung Cancer. *Cancer. Res.* **74**, 3259–3270 (2014).

(17) Schult, T. A. et al. Screening Human Lung Cancer with Predictive Models of Serum Magnetic Resonance Spectroscopy Metabolomics. *Proc. Natl. Acad. Sci. USA.* **118**, e2110633118 (2021).

(18) Shestakova, K. M. et al. Targeted Metabolomic Profiling as A Tool for Diagnostics of Patients with Non-Small-Cell Lung Cancer. *Sci. Rep.* **13**, 11072 (2023).

(19) Wen, T. et al. Exploratory Investigation of Plasma Metabolomics in Human Lung Adenocarcinoma. *Mol. Biosyst.* **9**, 2370–2378 (2013).

(20) Li, J. et al. Serum Untargeted Metabolomics Reveal Metabolic Alteration of Non-Small Cell Lung Cancer and Refine Disease Detection. *Cancer. Sci.* **114**, 680–689 (2023).

(21) Sun, T. et al. Lipidomics Reveals New Lipid-Based Lung Adenocarcinoma Early Diagnosis Model. EMBO. *Mol. Med.* **16**, 854–869 (2024).

(22) Nie, M.et al. Evolutionary Metabolic Landscape from Preneoplasia to Invasive Lung Adenocarcinoma. *Nat. Commun.* **12**, 6479 (2021).

(23) Wang, L. et al. Integrative Serum Metabolic Fingerprints Based Multi-Modal Platforms for Lung Adenocarcinoma Early Detection and Pulmonary Nodule Classification. *Adv. Sci.* **9**, e2203786 (2022).

(24) Yao, Y. et al. Metabolomic Differentiation of Benign vs Malignant Pulmonary Nodules with High Specificity via High-Resolution Mass Spectrometry Analysis of Patient Sera. *Nat. Commun.* **14**, 2339 (2023).

(25) Huang, L. et al. Machine Learning of Serum Metabolic Patterns Encodes Early-Stage Lung Adenocarcinoma. *Nat. Commun.* **11**, 3556 (2020).

(26) Zheng, R. et al. Machine Learning-Based Integrated Multiomics Characterization of Colorectal Cancer Reveals Distinctive Metabolic Signatures. *Anal. Chem.* **96**, 8772–8781 (2024).

(27) Odenkirk, M. T. et al. Multiomic Big Data Analysis Challenges: Increasing Confidence in the Interpretation of Artificial Intelligence Assessments. *Anal. Chem.* **93**, 7763–7773 (2021).

(28) Asef, C. K. et al. Unknown Metabolite Identification Using Machine Learning Collision Cross-Section Prediction and Tandem Mass Spectrometry. *Anal. Chem.* **95**, 1047–1056 (2023).

(29) Malta, T. M. et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell.* **173**, 338–354.e15 (2018).

(30) Konno, N. et al. Machine Learning Enables Prediction of Metabolic System Evolution in Bacteria. *Sci. Adv.* **9**, eadc9130 (2023).

(31) Greener, J. G. et al. A Guide To Machine Learning for Biologists. *Nat. Rev. Mol. Cell. Biol.* **23**, 40–55 (2022).

(32) Chen, J. et al. Machine Learning Aids Classification and Discrimination of Noncanonical DNA Folding Motifs by an Arrayed Host: Guest Sensing System. *J. Am. Chem. Soc.* **143**, 12791–12799 (2021).

(33) Chen, R. J. et al. Synthetic Data in Machine Learning for Medicine and Healthcare. *Nat. Biomed. Eng.* **5**, 493–497 (2021).

(34) Dong, X. et al. A Survey on Ensemble Learning. *Front. Comput. Sci.* **14**, 241–58 (2020).

(35) Heidari, B. M. et al. Culture-Free Identification and Metabolic Profiling of Microalgal Single Cells via Ensemble Learning of Ramanomes. *Anal. Chem.* **93**, 8872–8880 (2021).

(36) Janizek, J. D. et al. Uncovering Expression Signatures of Synergistic Drug Responses via Ensembles of Explainable Machine-Learning Models. *Nat. Biomed. Eng.* **7**, 811–829 (2023).

(37) Cao, Y. et al. Ensemble Deep Learning in Bioinformatics. *Nat. Mach. Intell.* **2**, 500–508 (2020).

(38) Arnaout, R. et al. An Ensemble of Neural Networks Provides Expert-Level Prenatal Detection of Complex Congenital Heart Disease. Nat. Med. **27**, 882–891 (2021).

(39) Canesin, G. et al. Heme-Derived Metabolic Signals Dictate Immune Responses. *Front. Immunol.* **11**, 66 (2020).

(40) Qian, X. et al. Integrated Microbiome, Metabolome, and Proteome Analysis Identifies a Novel Interplay Among Commensal Bacteria, Metabolites and Candidate Targets in Non-Small Cell

Lung Cancer. *Clin. Transl. Med.* **12**, e947 (2022).

(41) Lv, M. et al. Plasma Lipidomics Profiling to Identify the Biomarkers of Diagnosis and Radiotherapy Response for Advanced Non-Small-Cell Lung Cancer Patients. *J. Lipids.* **2024**, 6730504 (2024).

(42) Chen, L. et al. 25-Hydroxycholesterol Promotes Migration and Invasion of Lung Adenocarcinoma Cells. *Biochem. Biophys. Res. Commun.* **484**, 857–863 (2017).

(43) Musial, C. et al. Induction of 2-Hydroxycatecholestrogens O-Methylation: A Missing Puzzle Piece in Diagnostics and Treatment of Lung Cancer. *Redox. Biol.* **55**, 102395 (2022).

(44) Chen, X. et al. Whole-Lesion Computed Tomography-Based Entropy Parameters for the Differentiation of Minimally Invasive and Invasive Adenocarcinomas Appearing as Pulmonary Subsolid Nodules. *J. Comput. Assist. Tomogr.* **43**, 817–824 (2019).

(45) Zhang, J. et al. Why do Pathological Stage IA Lung Adenocarcinomas Vary From Prognosis?: A Clinicopathologic Study of 176 Patients with Pathological Stage IA Lung Adenocarcinoma Based on the IASLC/ATS/ERS Classification. *J. Thorac. Oncol.* **8**, 1196–202 (2013).

(46) Altorki, N. K. et al. Sublobar Resection Is Equivalent to Lobectomy for Clinical Stage 1A Lung Cancer in Solid Nodules. *J. Thorac. Cardiovasc. Surg.* **147**, 754–762; Discussion 762–764 (2014).

(47) Huang, Z. et al. From Purines to Purinergic Signalling: Molecular Functions and Human Diseases. *Signal. Transduct. Target. Ther.* **6**, 162 (2021).

(48) Ma, C. et al. Gut Microbiome-Mediated Bile Acid Metabolism Regulates Liver Cancer via NKT Cells. *Science.* **360**, eaan5931 (2018).

(49) Martin-Perez, M. et al. The Role of Lipids in Cancer Progression and Metastasis. *Cell. Metab.* **34**, 1675-1699 (2022).

(50) Nie, M. et al. Evolutionary Metabolic Landscape from Preneoplasia to Invasive Lung Adenocarcinoma. *Nat. Commun*. **12**, 6479 (2021).

(51) Arndt, M. A. et al. The Arginine Metabolite Agmatine Protects Mitochondrial Function and Confers Resistance to Cellular Apoptosis. *Am. J. Physiol. Cell. Physiol*. **296**, C1411-9 (2009).

(52) Dong, B. et al. Plasma Proteometabolome in Lung Cancer: Exploring Biomarkers through Bidirectional Mendelian Randomization and Colocalization Analysis. *Hum. Mol. Genet*. **33**, 1688-1696 (2024).

(53) Wang, C. et al. Multi-Omics Analyses Reveal Biological and Clinical Insights in Recurrent Stage I Non-Small Cell Lung Cancer. *Nat. Commun.* **16**, 1477 (2025).

(54) Zhang, Y. et al. Evolutionary Proteogenomic Landscape from Pre-Invasive to Invasive Lung Adenocarcinoma. *Cell. Rep. Med.* **5**, 101358 (2024).

(55) Nicholson, A. G. et al. The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015. *J. Thorac. Oncol.* **17**, 362-387 (2022).

(56) Li, R. et al. Deep Learning Applications in Computed Tomography Images for Pulmonary Nodule Detection and Diagnosis: A Review. *Diagnostics*. **12**, 298 (2022).

(57) Wieder, B. et al. Strong and Fragile Topological Dirac Semimetals with Higher-Order Fermi Arcs. *Nat. Commun.* **11**, 627 (2020).

(58) Chen, Y. C. et al. Multiomics Analysis Reveals Molecular Changes During Early Progression of Precancerous Lesions to Lung Adenocarcinoma in Never-Smokers. *Cancer. Res.* **85**, 602-

617 (2025).

**Fig. 1** Schematic illustration of lung adenocarcinoma diagnosis at very early stage with the ensemble machine learning-aided serum untargeted metabolomics.

**Fig. 2 Metabolic alteration during early carcinogenesis of LUAD.** (a) Clinical parameters of the study cohort. (b) Classes and counts of metabolites detected in blood serum of participants. (c) Sparse partial least squares discriminant analysis (sPLS-DA) of the metabolomics data from HC ($n$ = 50), AIS ($n$ = 50), MIA ($n$ = 49), and IAC ($n$ = 50) patients. (d) Volcano plot of significantly differential metabolites between IAC ($n$ = 50) and HC ($n$ = 50) groups, based on false discovery rate (FDR) < 0.05 and fold change (FC) > 1.25 or < 0.8. (e) KEGG (Kyoto Encyclopedia of Genes and Genomes) metabolic pathways enriched by significantly altered metabolites between the IAC ($n$ = 50) and the HC ($n$ = 50) groups.

**Fig. 3 Progressive evolution of serum metabolic landscape from HC to IAC.** (a) Heatmap of the relative abundance of the identified metabolites in HC ($n$ = 50), AIS ($n$ = 50), MIA ($n$ = 49), and IAC ($n$ = 50) groups. (b) Mfuzz clustering of the autoscaled intensity data of differential metabolites across four pathological stages (HC, AIS, MIA, and IAC). Six distinct temporal trends (Trends 1-6) represent characteristic patterns of metabolic alterations during LUAD progression, with positive and negative values indicating relative increases or decreases compared with the overall mean abundance. (c) Relative abundance distribution of metabolites in Trends 3 and 5 (one-way ANOVA). Dashed lines indicate the median, and short-dashed lines indicate 25th–75th percentiles.

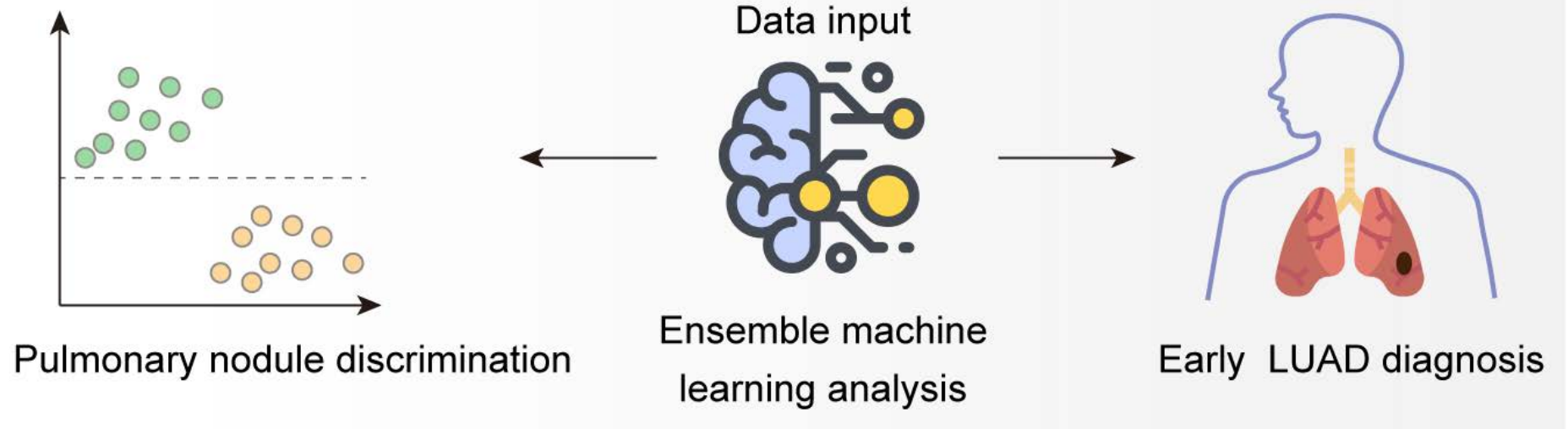
**Fig. 4 Machine learning analysis of circulating serum metabolites for non-invasive diagnosis of early IAC.** (a) Receiver operating characteristic (ROC) curve and (b) confusion matrix generated from the prediction of test samples (HC, $n$ = 10; IAC, $n$ = 10). (c) Relative abundance of individual metabolites in HC ($n$ = 40) and IAC ($n$ = 40) serum

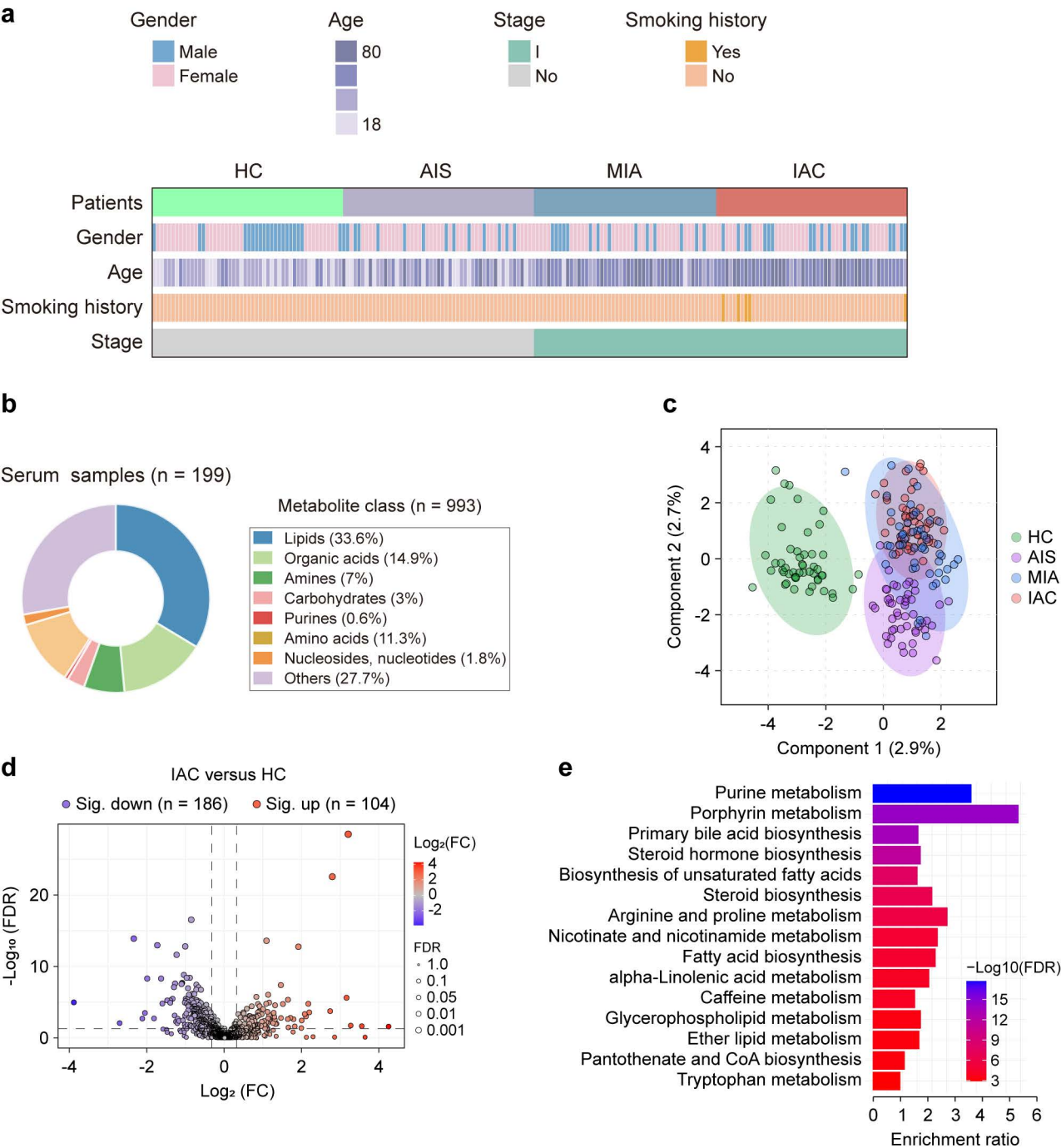
(two-sided Welch's t-test). Dashed line indicates the median; short-dashed lines indicate the 25th–75th percentiles. (d) Performance metrics of the six-metabolite classification model on the discovery and test sets. (e) SHapley Additive exPlanations (SHAP) analysis showing the contribution of each metabolite to the model's predictive performance. (f) Predictive performance of the six-metabolite model for distinguishing HC (green) from early IAC (purple) on the test set. The dotted line indicates the cutoff value of 0.5 used to separate predicted HC (lower right) from IAC (upper left).

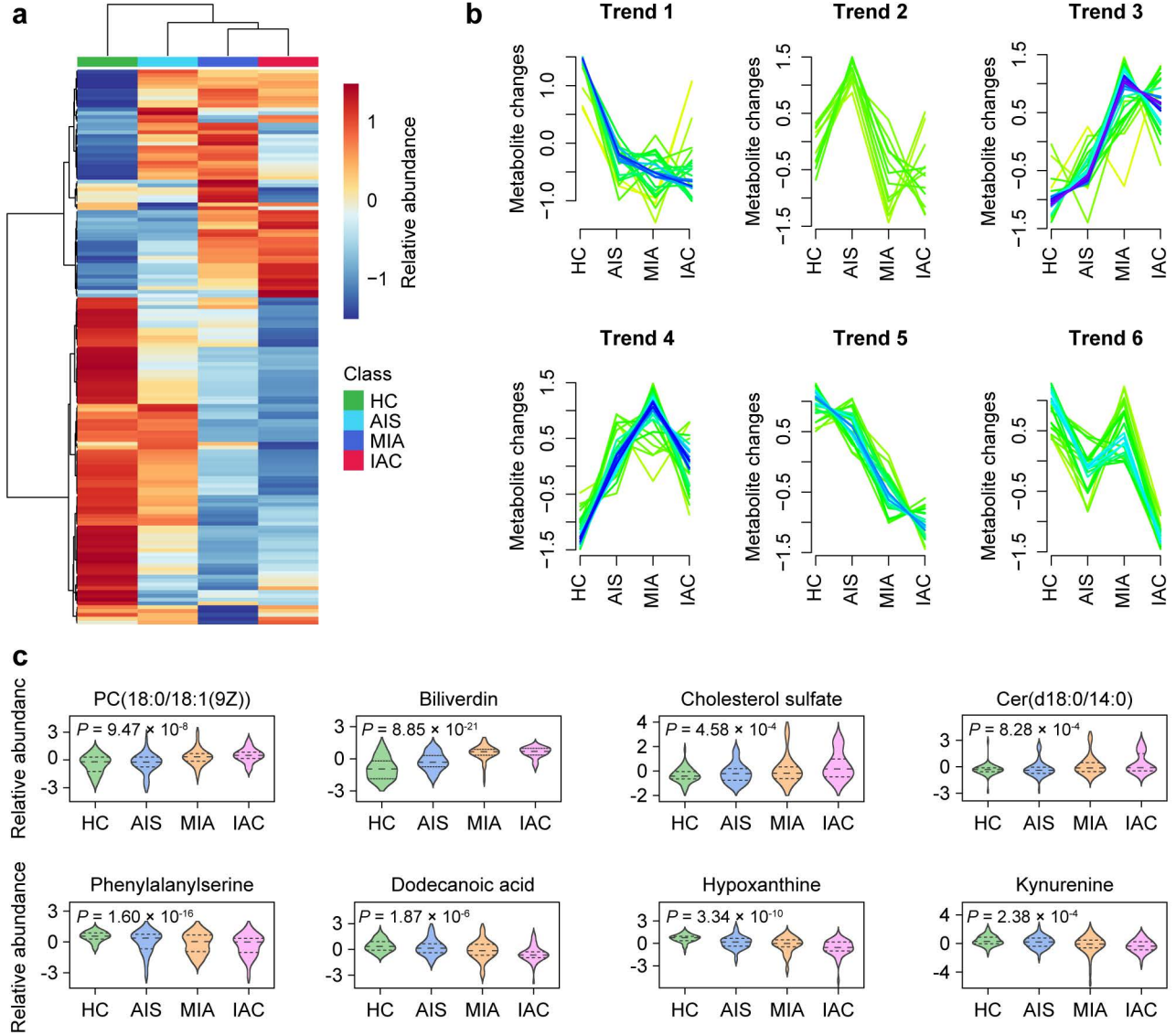
**Fig. 5 Machine learning analysis of circulating serum metabolites for discrimination of HC from AIS + MIA + IAC.** (a) Receiver operating characteristic (ROC) curve and (b) confusion matrix generated from the prediction of test samples (HC, *n* = 10; combined AIS + MIA + IAC, *n* = 30). (c) Relative abundance of individual metabolites in HC (*n* = 40) and combined AIS + MIA + IAC (*n* = 119) serum (two-sided Welch's t-test). Circle represents the median; short-dashed lines indicate the 25th–75th percentiles. (d) Performance metrics of the six-metabolite classification model on the discovery and test sets. (e) SHapley Additive exPlanations (SHAP) analysis showing the contribution of each metabolite to the model's predictive performance. (f) Predictive performance of the six-metabolite model for distinguishing HC (green) from AIS + MIA + IAC (purple) on the test set. The dotted line indicates the cutoff value of 0.5 used to separate predicted HC (lower right) from LUAD (upper left).

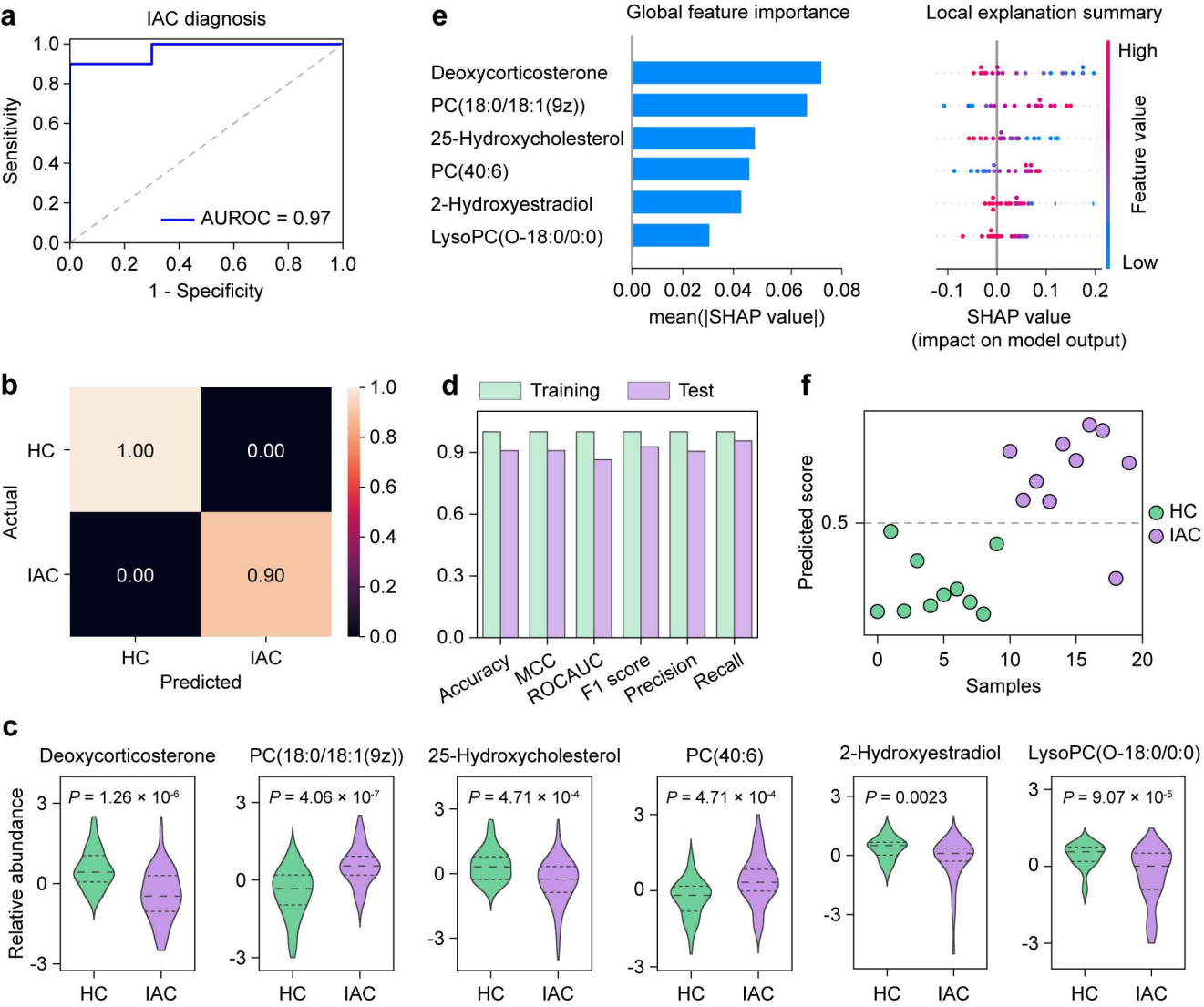
**Fig. 6 Machine learning analysis of circulating serum metabolites for discrimination of AIS from MIA + IAC.** (a) Receiver operating characteristic (ROC) curve and (b) confusion matrix generated from the prediction of test samples (HC, *n* = 10; MIA + IAC, *n* = 20). (c) Relative abundance of individual metabolites in HC (*n* = 40) and IAC (*n* = 79) serum (two-sided Welch's t-test). Circle represents the media; short -dashed lines indicate the 25th–75th percentiles. (d) SHapley Additive exPlanations (SHAP) analysis showing the contribution of each metabolite to the model's predictive performance.

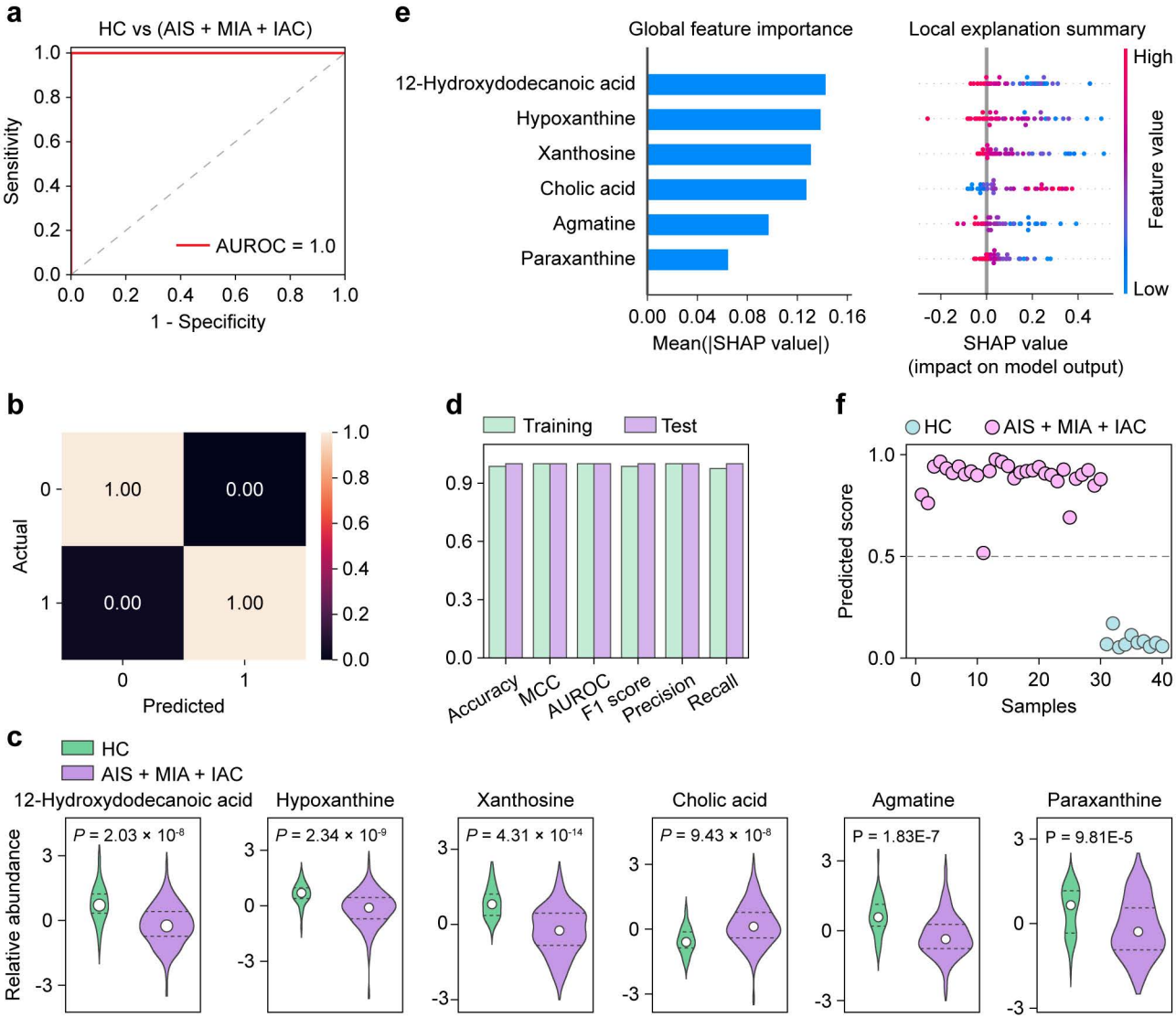
**Fig. 7 Differentiation between MIA and IAC.** (a) Receiver operating characteristic (ROC) curve. (b) Relative abundance of beta-glycerophosphoric acid and cytidine in MIA ($n$ = 39) and IAC ($n$ = 40) serum (two-sided Welch's t-test). Solid lines indicate the median; boxes show the 25th–75th percentiles; whiskers indicate 1.5 × interquartile range (IQR).

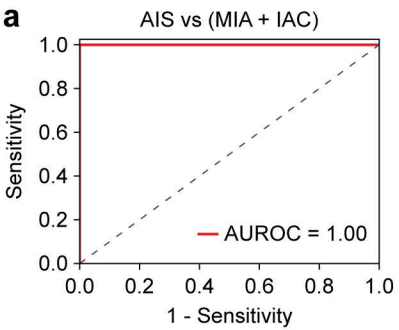



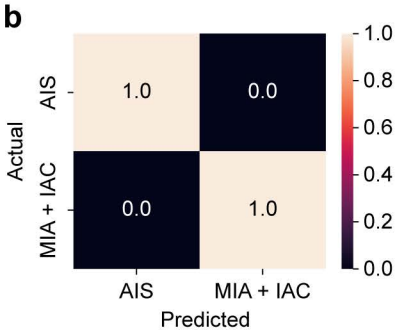

a
Gender
Male
Female
Age
80
18
Stage
I
No
Smoking history
Yes
No
HC
AIS
MIA
IAC
Patients
Gender
Age
Smoking history
Stage
b
Serum samples (n = 199)
Metabolite class (n = 993)
Lipids (33.6%)
Organic acids (14.9%)
Amines (7%)
Carbohydrates (3%)
Purines (0.6%)
Amino acids (11.3%)
Nucleosides, nucleotides (1.8%)
Others (27.7%)
c
Component 2 (2.7%)
Component 1 (2.9%)
HC
AIS
MIA
IAC
d
IAC versus HC
Sig. down (n = 186)
Sig. up (n = 104)
-Log10 (FDR)
Log2 (FC)
Log2(FC)
FDR
1.0
0.1
0.05
0.01
0.001
e
Purine metabolism
Porphyrin metabolism
Primary bile acid biosynthesis
Steroid hormone biosynthesis
Biosynthesis of unsaturated fatty acids
Steroid biosynthesis
Arginine and proline metabolism
Nicotinate and nicotinamide metabolism
Fatty acid biosynthesis
alpha-Linolenic acid metabolism
Caffeine metabolism
Glycerophospholipid metabolism
Ether lipid metabolism
Pantothenate and CoA biosynthesis
Tryptophan metabolism
−Log10(FDR)
Enrichment ratio

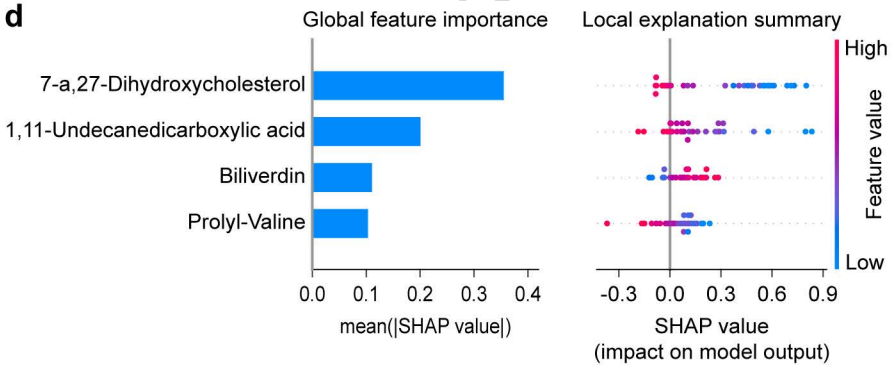

a
1
0
−1
Relative abundance
Class
HC
AIS
MIA
IAC
b
Trend 1
Trend 2
Trend 3
Trend 4
Trend 5
Trend 6
Metabolite changes
HC
AIS
MIA
IAC
c
Relative abundanc
PC(18:0/18:1(9Z))
P = 9.47 × 10-8
Biliverdin
P = 8.85 × 10-21
Cholesterol sulfate
P = 4.58 × 10-4
Cer(d18:0/14:0)
P = 8.28 × 10-4
Relative abundance
Phenylalanylserine
P = 1.60 × 10-16
Dodecanoic acid
P = 1.87 × 10-6
Hypoxanthine
P = 3.34 × 10-10
Kynurenine
P = 2.38 × 10-4

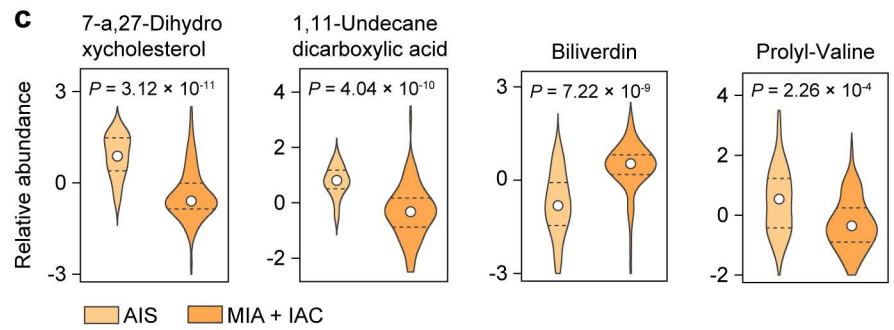

a
IAC diagnosis
Sensitivity
1 - Specificity
AUROC = 0.97
b
HC
IAC
Actual
Predicted
1.00
0.00
0.00
0.90
c
Deoxycorticosterone
PC(18:0/18:1(9z))
25-Hydroxycholesterol
PC(40:6)
2-Hydroxyestradiol
LysoPC(O-18:0/0:0)
Relative abundance
P = 1.26 × 10-6
P = 4.06 × 10-7
P = 4.71 × 10-4
P = 4.71 × 10-4
P = 0.0023
P = 9.07 × 10-5
d
Training
Test
Accuracy
MCC
ROCAUC
F1 score
Precision
Recall
e
Global feature importance
Deoxycorticosterone
PC(18:0/18:1(9z))
25-Hydroxycholesterol
PC(40:6)
2-Hydroxyestradiol
LysoPC(O-18:0/0:0)
mean(|SHAP value|)
Local explanation summary
High
Feature value
Low
SHAP value
(impact on model output)
f
Predicted score
Samples
HC
IAC


a
HC vs (AIS + MIA + IAC)
Sensitivity
1 - Specificity
AUROC = 1.0
e
Global feature importance
12-Hydroxydodecanoic acid
Hypoxanthine
Xanthosine
Cholic acid
Agmatine
Paraxanthine
Mean(|SHAP value|)
Local explanation summary
High
Feature value
Low
SHAP value
(impact on model output)
b
Actual
Predicted
1.00
0.00
0.00
1.00
d
Training
Test
Accuracy
MCC
AUROC
F1 score
Precision
Recall
f
HC
AIS + MIA + IAC
Predicted score
Samples
c
HC
AIS + MIA + IAC
12-Hydroxydodecanoic acid
P = 2.03 × 10-8
Hypoxanthine
P = 2.34 × 10-9
Xanthosine
P = 4.31 × 10-14
Cholic acid
P = 9.43 × 10-8
Agmatine
P = 1.83E-7
Paraxanthine
P = 9.81E-5
Relative abundance


a
AIS vs (MIA + IAC)
Sensitivity
1 - Sensitivity
AUROC = 1.00
0.0
0.2
0.4
0.6
0.8
1.0


b
Actual
Predicted
AIS
MIA + IAC
1.0
0.0
0.0
1.0


d
Global feature importance
7-a,27-Dihydroxycholesterol
1,11-Undecanedicarboxylic acid
Biliverdin
Prolyl-Valine
mean(|SHAP value|)
Local explanation summary
High
Feature value
Low
SHAP value
(impact on model output)


c
7-a,27-Dihydro
xycholesterol
1,11-Undecane
dicarboxylic acid
Biliverdin
Prolyl-Valine
Relative abundance
P = 3.12 × 10-11
P = 4.04 × 10-10
P = 7.22 × 10-9
P = 2.26 × 10-4
AIS
MIA + IAC