High Throughput Recurrent Pregnancy Loss Screening: Urine Metabolic Fingerprints via LDI-MS and Machine Learning

Yijiao Qu[b,c#], Ming Chen[a,d#], Mufeng Han[e], Xiaoyu Yu[f], Xi Yu[b,c], Jinghan Fan[b,c], Huihui Liu[b,c*], Liping Wang[a*], Zongxiu Nie[b,c*]

[a]*Centre of Reproductive Medicine, Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University, Shenzhen 518000, China*

[b]*Beijing National Laboratory for Molecular Sciences, Key Laboratory of Analytical Chemistry for Living Biosystems, Institute of Chemistry, Chinese Academy of Sciences, Beijing, 100190, China*

[c]*University of Chinese Academy of Sciences, Beijing, 100190, China*

[d]*Department of Gynecology and Obstetrics, Guangxi University of Chinese Medicine, Nanning, 530200, China*

[e]*Beijing National Day School, Beijing 100039, China*

[f]*Peking University Health Science Center, Beijing, 100191, China*

 # These authors contributed equally to this work.

Corresponding author：

**Huihui Liu Email: hhliu@iccas.ac.cn*

**Liping Wang Email: wlilyu@hotmail.com.*

**Zongxiu Nie Email: znie@iccas.ac.cn*

Methods

Machine learning method

Machine learning analysis was conducted with Orange (Version 3.38.1). in Python 3.10. The build in classifier logistic regression (LR), gradient boosting (GB), neural network (NN), random forest (RF) and support vector machine (SVM) were applied. Model parameters were set as follows:

LR in distinguishing RPL and HC: Regularization: Lasso (L1), C=1, class weights=False.

GB in distinguishing RPL and HC: Method: Gradient Boosting (scikit-learn), Number of trees: 100, Learning rate: 0.100, Replicable training: Ture, Limit depth of individual trees: 3, Do not split subsets smaller than: 2, Fraction of training instances: 1.00.

NN in distinguishing RPL and HC: Hidden layers: 100, Activation: ReLu, Solver: Adam, Alpha: 0.0001, Max iterations: 90, Replicable training: True.

RF in distinguishing RPL and HC: Number of trees: 25, Number of attributes considered at each split: False, Limit depth of individual trees: 6, Do not split subsets smaller than: 12.

SVM in distinguishing RPL and HC: SVM type: v-SVM, C=1.00, v=0.50, Kernel: RBF, $\exp(-\text{auto}|x-y|^2)$, Numerical tolerance: 0.0010, Iteration limit: 100.

Figures

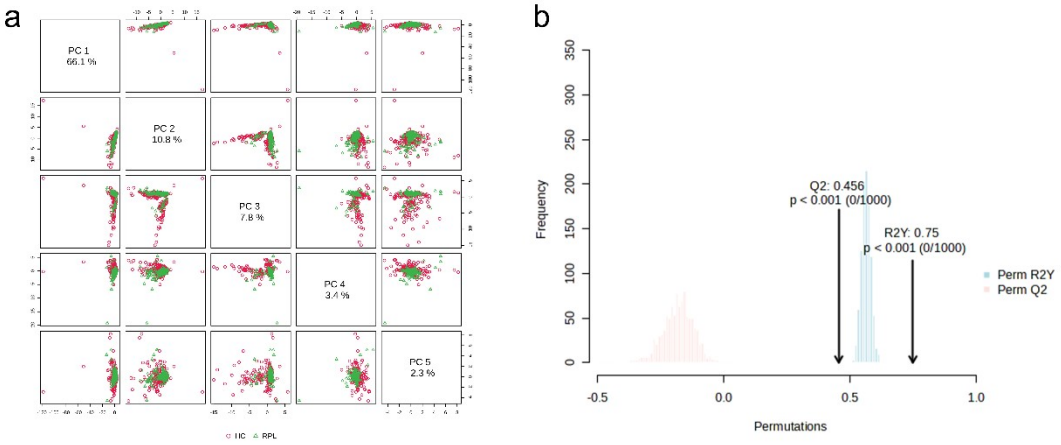

Figure S1. (a) Overview of PCA results in the classification of Recurrent Pregnancy Loss (RPL) and healthy control (HC) group. (b) Permutation test results and validation information for the OPLS-DA model of RPL and HC.

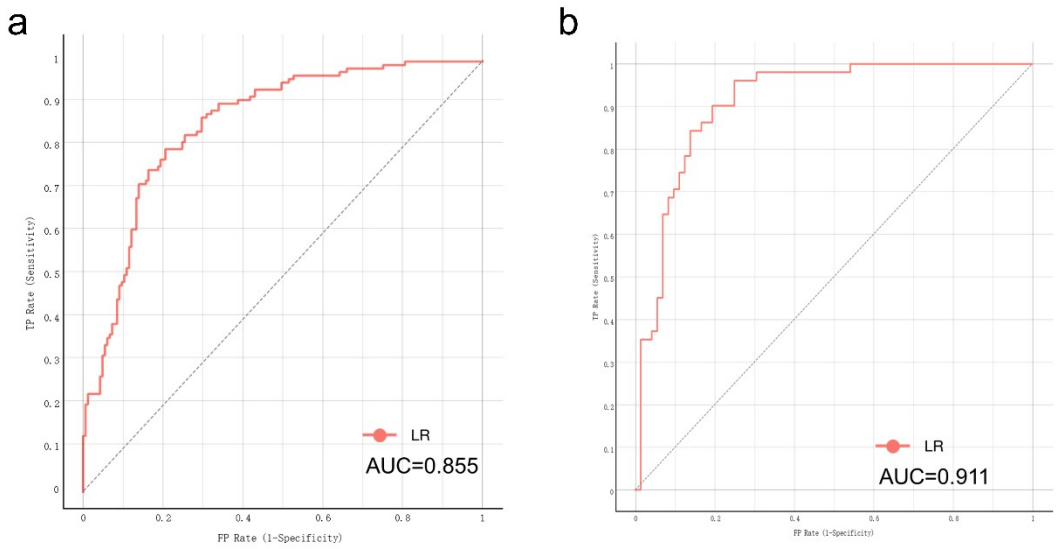

Figure S2. ROC curve of RPL distinguishment with logistic regression (LR) in training set (a) and in test set (b).

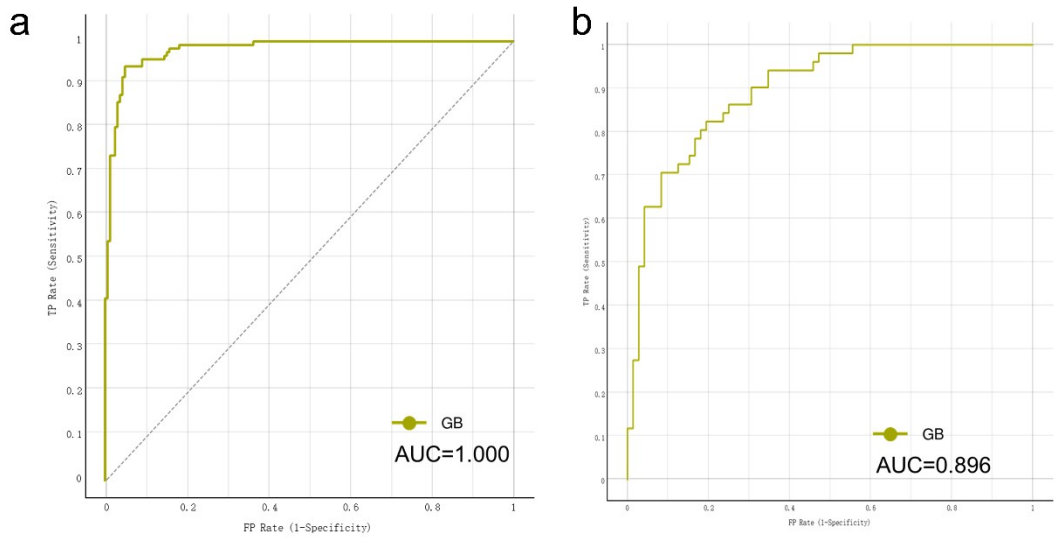

Figure S3. ROC curve of RPL distinguishment with gradient boosting (GB) in training set (a) and in test set (b).

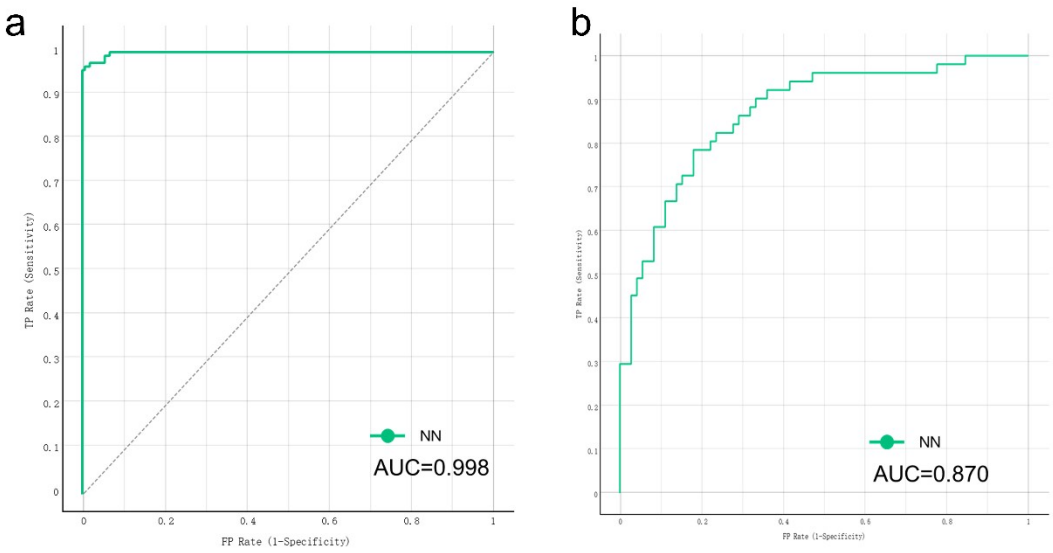

Figure S4. ROC curve of RPL distinguishment with neural network (NN) in training set (a) and in test set (b).

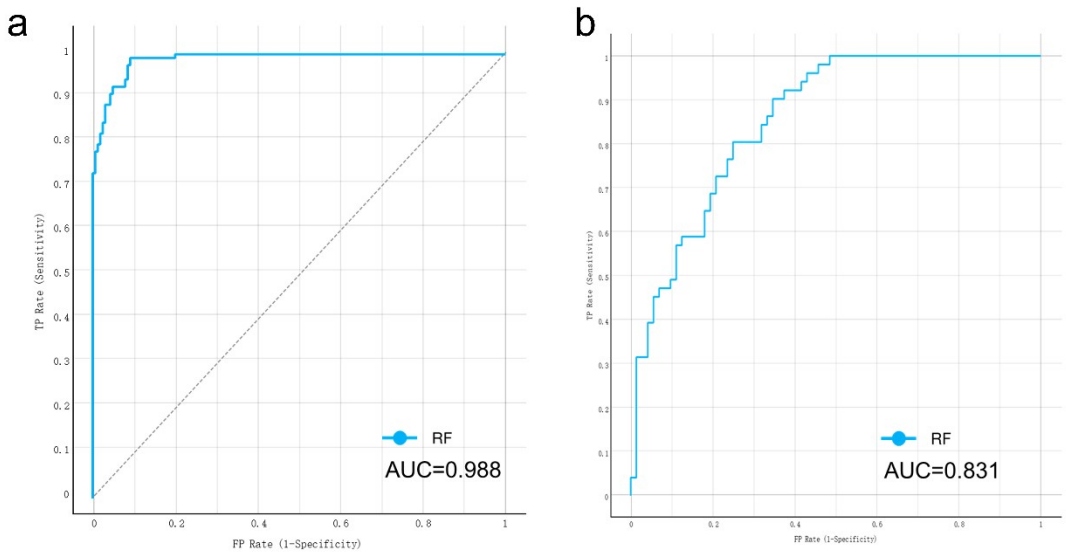

Figure S5. ROC curve of RPL distinguishment with random forest (RF) in training set (a) and in test set (b).

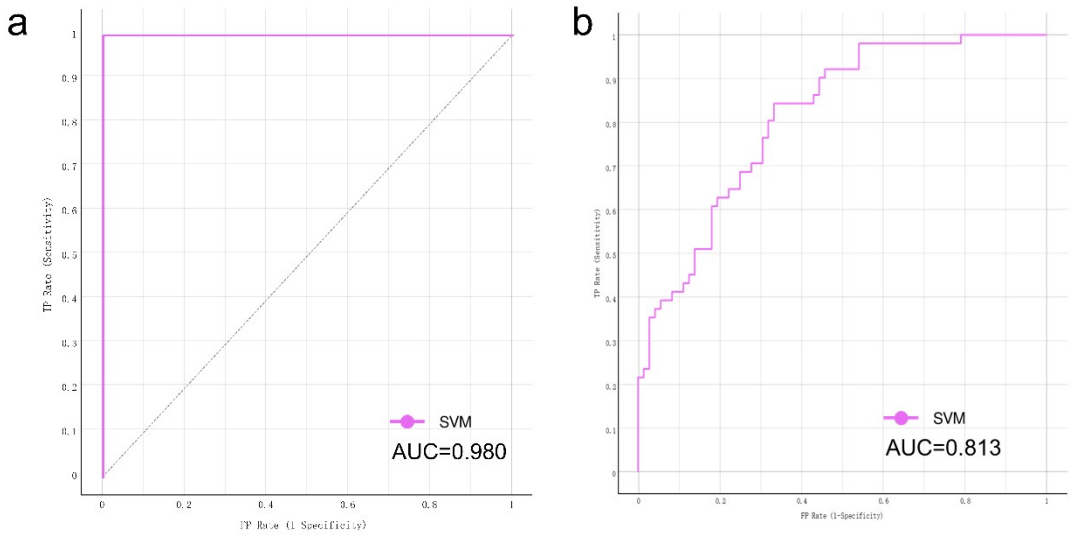

Figure S6. ROC curve of RPL distinguishment with support vector machine (SVM) in training set (a) and in test set (b).

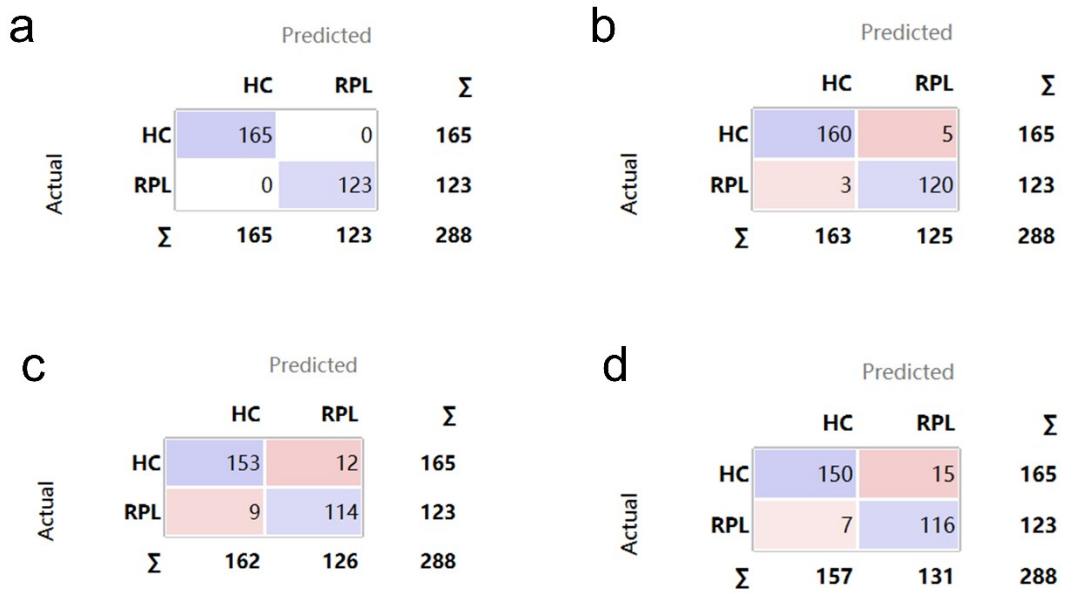

Figure S7. (a) Confusion matrix of the training set with GB. (b) Confusion matrix of the training set with NN. (c) Confusion matrix of the training set with RF. (d) Confusion matrix of the training set with SVM.

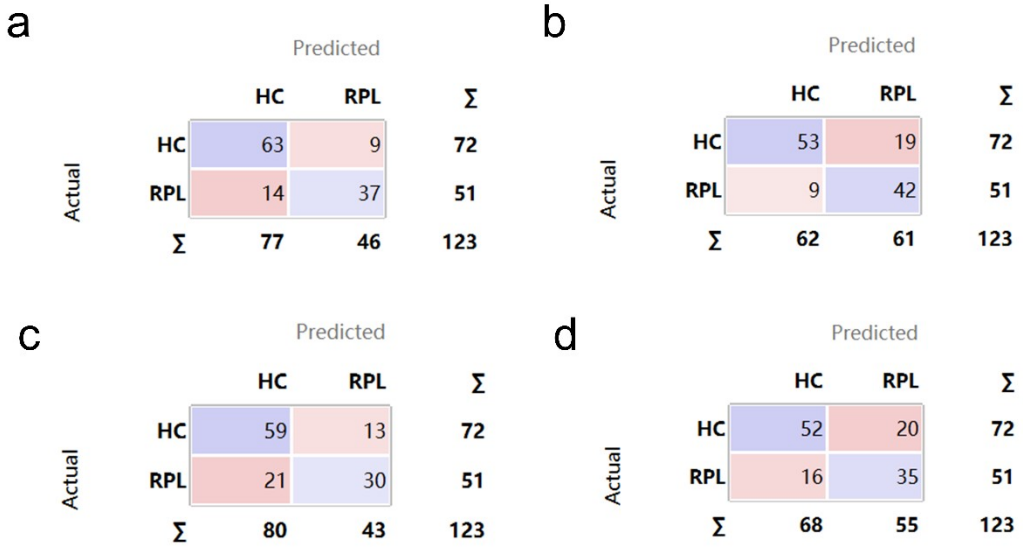

Figure S8. (a) Confusion matrix of the test set with GB. (b) Confusion matrix of the test set with NN. (c) Confusion matrix of the test set with RF. (d) Confusion matrix of the test set with SVM.

Table. S1 The urine metabolites identified by AP-SMALDI MS in positive ion mode.

Name	Formula	Adduct	Theoretical	Experimental	Δppm
Creatinine	$C_4H_7N_3O$	M+K	152.0221	152.0220	0.62
Proline betaine	$C_7H_{13}NO_2$	M+K	182.0578	182.0576	1.29
Alanylglycine	$C_5H_{10}N_2O_3$	M+2Na-H	191.0403	191.0402	0.65
Taurine	$C_2H_7NO_3S$	M+2K-H	201.9337	201.9336	0.44
4-Hydroxy-benzenepropanedioate	$C_9H_8O_5$	M+Na	219.0264	219.0267	-1.25
Glucose	$C_6H_{12}O_6$	M+K	219.0265	219.0267	-0.75
5-Hydroxyindoleacetaldehyde	$C_{10}H_9NO_2$	M+2Na-H	220.0345	220.0340	2.16
Galactitol	$C_6H_{14}O_6$	M+K	221.0422	221.0424	-0.95
L-Glutamine	$C_5H_{10}N_2O_3$	M+2K-H	222.9882	222.9882	0.13
N-(3-acetamidopropyl)pyrrolidin-2-one	$C_9H_{16}N_2O_2$	M+K	223.0843	223.0841	1.00
2-Amino-4-hydroxy-6-pteridinecarboxylic acid	$C_7H_5N_5O_3$	M+Na	230.0285	230.0286	-0.50
L-Histidine	$C_6H_9N_3O_2$	M+2K-H	231.9885	231.9884	0.59
Isoetharine	$C_{13}H_{21}NO_3$	M+H	240.1594	240.1616	-9.11
N-Acetylserotonin	$C_{12}H_{14}N_2O_2$	M+Na	241.0947	241.0947	0.28
Dechloroethylifosfamide	$C_5H_{12}ClN_2O_2P$	M+2Na-H	243.0037	243.0034	1.28
Uric acid	$C_5H_4N_4O_3$	M+2K-H	244.9474	244.9475	-0.54

Methylhistidine	$C_7H_{11}N_3O_2$	M+2K-H	246.0042	246.0047	-2.33
Symmetric dimethylarginine	$C_8H_{18}N_4O_2$	M+2Na-H	247.1141	247.1143	-0.59
L-Tryptophan	$C_{11}H_{12}N_2O_2$	M+2Na-H	249.0610	249.0609	0.72
Phenol sulphate	$C_6H_6O_4S$	M+2K-H	250.9177	250.9172	2.09
Glucosamine	$C_6H_{13}NO_5$	M+2K-H	255.9984	255.9979	1.99
Hippuric acid	$C_9H_9NO_3$	M+2K-H	255.9773	255.9770	1.07
L-Aspartyl-4-phosphate	$C_4H_8NO_7P$	M+2Na-H	257.9750	257.9747	1.02
3-(3-Hydroxyphenyl)-3-hydroxypropanoic acid	$C_9H_{10}O_4$	M+2K-H	258.9769	258.9774	-1.83
N-Acetylmannosamine	$C_8H_{15}NO_6$	M+K	260.0531	260.0526	2.08
N-Despyridinyl rosiglitazone	$C_{13}H_{16}N_2O_3S$	M+H-H_2O	263.0855	263.0854	0.46
Vanilloylglycine	$C_{10}H_{11}NO_5$	M+K	264.0269	264.0266	1.22
N6-Acetyl-L-lysine	$C_8H_{16}N_2O_3$	M+2K-H	265.0351	265.0349	0.74
p-Cresol sulfate	$C_7H_8O_4S$	M+2K-H	264.9334	264.9340	-2.26
Thymidine	$C_{10}H_{14}N_2O_5$	M+Na	265.0795	265.0790	1.97
Leucylproline	$C_{11}H_{20}N_2O_3$	M+K	267.1106	267.1110	-1.76
5,6-Dihydrouridine	$C_9H_{14}N_2O_6$	M+Na	269.0744	269.0738	2.26
Phosphocreatinine	$C_4H_8N_3O_4P$	M+2K-H	269.9443	269.9448	-1.86

gamma-Glutamylthreonine	$C_9H_{16}N_2O_6$	M+Na	271.0901	271.0904	-1.35
3-Hydroxyhippuric acid	$C_9H_9NO_4$	M+2K-H	271.9722	271.9726	-1.45
5-Acetylamino-6-amino-3-methyluracil	$C_7H_{10}N_4O_3$	M+2K-H	274.9943	274.9937	2.12
O-Phosphothreonine	$C_4H_{10}NO_6P$	M+2K-H	275.9436	275.9446	-3.49
Indole-3-acetylglycine	$C_{12}H_{12}N_2O_3$	M+2Na-H	277.0560	277.0567	-2.39
Methotrimeprazine	$C_{19}H_{24}N_2OS$	M+H-2H_2O	293.1482	293.1478	1.51
Phenol glucuronide	$C_{12}H_{14}O_7$	M+Na	293.0632	293.0634	-0.77
Anisindione	$C_{16}H_{12}O_3$	M+2Na-H	297.0498	297.0498	0.17
Creatine riboside	$C_9H_{17}N_3O_6$	M+K	302.0749	302.0783	-11.40
Carnosine	$C_9H_{14}N_4O_3$	M+2K-H	303.0256	303.0259	-1.04
1-Methyladenosine	$C_{11}H_{15}N_5O_4$	M+Na	304.1016	304.1020	-1.46
Demonomethylchlorpromazine	$C_{16}H_{17}ClN_2S$	M+H	305.0874	305.0879	-1.77
Prolyl-Hydroxyproline	$C_{10}H_{16}N_2O_4$	M+2K-H	305.0300	305.0304	-1.24
Cinoxacin	$C_{12}H_{10}N_2O_5$	M+2Na-H	307.0301	307.0306	-1.56
Dihydroformononetin	$C_{16}H_{14}O_4$	M+K	309.0524	309.0505	6.10
N-Acetylcystathionine	$C_9H_{16}N_2O_5S$	M+2Na-H	309.0492	309.0489	1.02
N-Desmethylpromazine	$C_{16}H_{18}N_2S$	M+K	309.0822	309.0815	2.27

Diisobutyl phthalate	$C_{16}H_{22}O_4$	M+K	317.1150	317.1142	2.39
Monoethylhexyl phthalic acid	$C_{16}H_{22}O_4$	M+K	317.1150	317.1142	2.49
5'-Methylthioadenosine	$C_{11}H_{15}N_5O_3S$	M+Na	320.0788	320.0785	1.06
1-Methylinosine	$C_{11}H_{14}N_4O_5$	M+K	321.0596	321.0602	-2.09
Caffeic acid O-glucuronide	$C_{15}H_{16}O_{10}$	M+H-2H_2O	321.0616	321.0620	-1.16
Pseudouridine	$C_9H_{12}N_2O_6$	M+2K-H	320.9886	320.9878	2.57
Uridine	$C_9H_{12}N_2O_6$	M+2K-H	320.9886	320.9878	2.64
N-acetyltryptophan	$C_{13}H_{14}N_2O_3$	M+2K-H	323.0195	323.0200	-1.54
Xanthosine	$C_{10}H_{12}N_4O_6$	M+K	323.0388	323.0392	-1.00
2,2,4-Trimethyl-1,3-pentadienol diisobutyrate	$C_{16}H_{30}O_4$	M+K	325.1776	325.1772	1.34
Pentahydroxyisoflavone	$C_{15}H_{10}O_7$	M+Na	325.0319	325.0329	-3.12
O-Desmethylindomethacin	$C_{18}H_{14}ClNO_4$	M+H-H_2O	326.0584	326.0588	-1.24
gallocatechin	$C_{15}H_{14}O_7$	M+Na	329.0632	329.0637	-1.46
Hydroxytyrosol 3'-glucuronide	$C_{14}H_{18}O_9$	M+H	331.1024	331.1027	-0.96
Chlorothiazide	$C_7H_6ClN_3O_4S_2$	M+K	333.9120	333.9120	-0.04
Didemethylcitalopram	$C_{18}H_{17}FN_2O$	M+K	335.0956	335.0962	-1.80
1-Methylguanosine	$C_{11}H_{15}N_5O_5$	M+K	336.0705	336.0703	0.58

Captopril-cysteine disulfide	$C_{12}H_{20}N_2O_5S_2$	M+H	337.0886	337.0895	-2.58
Phenylacetylglutamine	$C_{13}H_{16}N_2O_4$	M+2K-H	341.0300	341.0295	1.49
N2,N2-Dimethylguanosine	$C_{12}H_{17}N_5O_5$	M+K	350.0861	350.0869	-2.14
Cyclic AMP	$C_{10}H_{12}N_5O_6P$	M+Na	352.0417	352.0411	1.84
Urolithin B 3-O-glucuronide	$C_{19}H_{16}O_9$	M+H-2H2O	353.0667	353.0661	1.67
Aspartylphenylalanine	$C_{13}H_{16}N_2O_5$	M+2K-H	357.0250	357.0251	-0.49
Tyrosol glucuronide	$C_{14}H_{18}O_8$	M+2Na-H	359.0713	359.0705	2.35
Argininosuccinic acid	$C_{10}H_{18}N_4O_6$	M+2K-H	367.0417	367.0416	0.22
3'-O-Methyl-(-)-epicatechin-5-O-sulphate	$C_{16}H_{16}O_7S$	M+Na	375.0509	375.0500	2.49
Neomenthol-glucuronide	$C_{16}H_{28}O_7$	M+2Na-H	377.1547	377.1556	-2.59
D-Maltose	$C_{12}H_{22}O_{11}$	M+K	381.0794	381.0788	1.61
N-Desmethyl-hydroxy rosiglitazone	$C_{17}H_{17}N_3O_4S$	M+Na	382.0832	382.0831	0.24
beta-1,4-Mannosyl-N-acetylglucosamine	$C_{14}H_{25}NO_{11}$	M+K	422.1059	422.1048	2.54
Sulfoxone	$C_{14}H_{16}N_2O_6S_3$	M+K	442.9802	442.9788	3.27
Glucosylgalactosyl hydroxylysine	$C_{18}H_{34}N_2O_{13}$	M+K	525.1692	525.1694	-0.33

Table. S2 Age distribution of study participants.

Groups	Number	Age Distribution		
		20-29	30-39	40-50
Healthy Control	78	9	50	19
Recurrent Pregnancy Loss	58	17	37	4

Table. S3 Metrics of classifiers for RPL versus HC in training set.

Model	AUC	CA	F1	Prec	Spec
Logistic Regression	0.855	0.792	0.792	0.795	0.793
Gradient Boosting	1.000	1.000	1.000	1.000	1.000
Neural Network	0.998	0.972	0.972	0.972	0.973
Random Forest	0.988	0.927	0.927	0.927	0.927
SVM	0.980	0.924	0.924	0.926	0.929

Table. S4 Metrics of classifiers for RPL versus HC in test set.

Model	AUC	CA	F1	Prec	Spec
Logistic Regression	0.911	0.846	0.846	0.848	0.845
Gradient Boosting	0.896	0.813	0.811	0.812	0.787
Neural Network	0.870	0.772	0.774	0.786	0.787
Random Forest	0.831	0.707	0.705	0.705	0.678
SVM	0.813	0.707	0.709	0.711	0.701