Supporting information

Endometriosis detection using blood plasma through data fusion of Fourier transform infrared spectroscopy and high-performance liquid chromatography.

Sanoop Pavithran m[1], Clint Mathew[1], Rekha Upadhya[2], Anjali M[3], Jacinta S. D'Souza[4], Muraleedhar V pai[5], Santhosh Chidangil[1*]

1 Manipal Institute of Applied Physics, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India.

2 Department of Obstetrics and Gynaecology, Kasturba Medical College, Manipal, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India.

3 Department of Reproductive Medicine and Surgery, Kasturba Medical College, Manipal, Manipal Academy of Higher Education, Manipal, Karnataka, India.

4 School of Biological Sciences, University of Mumbai-Department of Atomic Energy Centre for Excellence in Basic Sciences, Mumbai, Maharashtra, India.

5 Department of Obstetrics and Gynaecology, Sikkim Manipal Institute of Medical Sciences, Sikkim Manipal University, Gangtok, Sikkim, India.

*Corresponding author : Santhosh Chidangil(Santhosh.cls@manipal.edu)

Supporting information

Feature Selection and data Fusion of Fourier transform infrared spectroscopy and high-performance liquid chromatography of blood plasma for Endometriosis detection.

S1

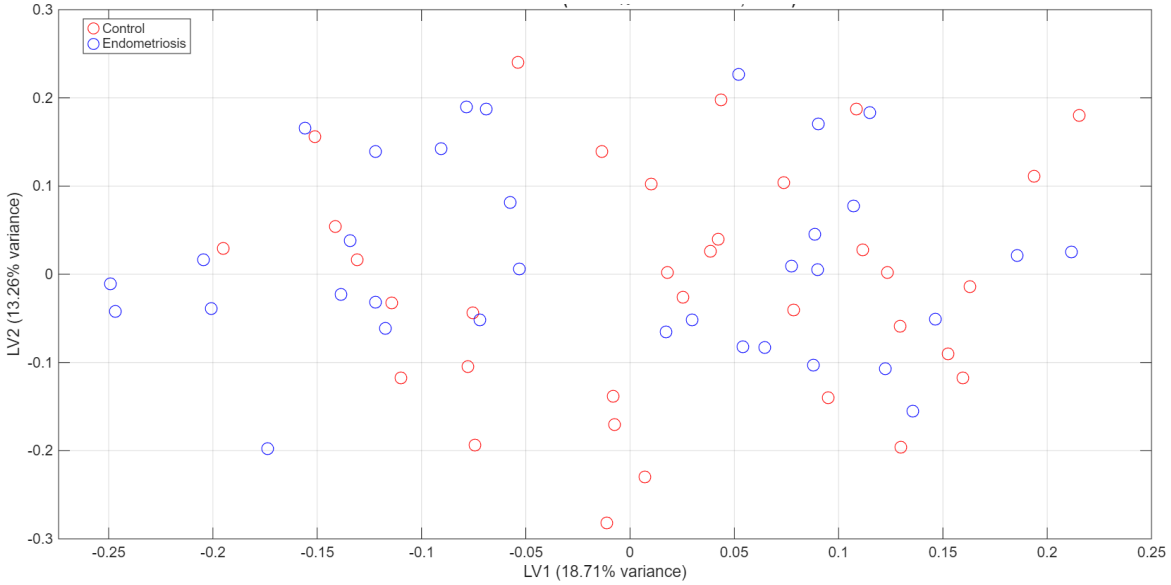

Figure 1: The PLS-DA score plot of ATR-FTIR data.

S2

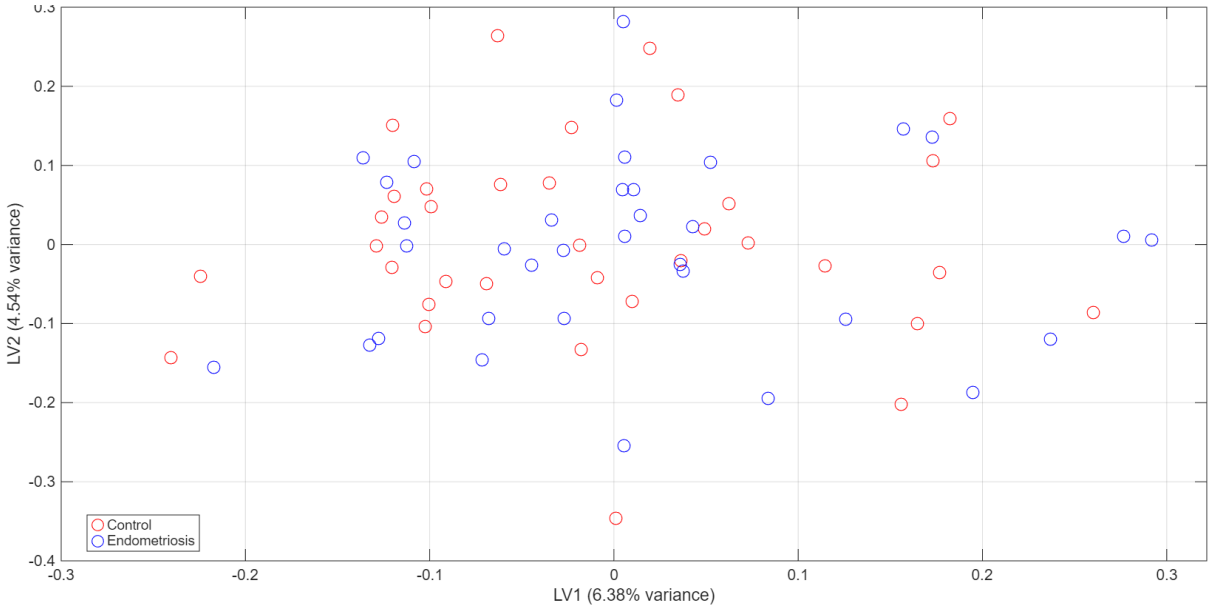

Figure 1: The PLS-DA score plot of HPLC-FLD data.

Table S1

Train data: ATR-FTIR

Data for classification	SVM		KNN		ANN	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Whole data	48.3	54.3	31	82.9	55.2	54.3
GA selected data	51.7	74.3	27.6	77.1	69	71.4
VIP selected data	55.2	45.7	44.8	42.9	62.1	62.9
RFIV selected data	64.3	66.7	60.7	72.2	50	66.7
Common selected data among any of 2 or more FS algorithms	24.1	65.7	17.2	80	48.3	65.7

Train data: HPLC-FLD

Data for classification						
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Whole data	76.7	72.2	46.7	83.3	60	69.4
GA selected data	80	80.6	53.3	75	66.7	75
VIP selected data	86.7	58.3	50	72.2	63.3	66.7
RFIV selected data	80	58.3	70	63.9	40	69.4
Common selected data among any of 2 or more FS algorithms	63.3	77.8	36.7	75	53.3	80.6

Table S2: Comparison of evaluation of ML model on feature selected concatenated data.

Fused data for classification	SVM				KNN				ANN			
	Accuracy	Sensitivity	Specificity	AUC of ROC	Accuracy	Sensitivity	Specificity	AUC of ROC	Accuracy	Sensitivity	Specificity	AUC of ROC
LLDF with feature selection GA	64.3	83.3	50	0.73	78.6	66.7	62.5	0.66	75	100	50	0.72
LLDF with feature selection VIP	64.3	83.3	50	0.77	71.4	83.3	62.5	0.71	71.4	100	50	0.81
LLDF with feature selection RFIV	71.4	100	50	0.71	71.4	100	50	0.77	71.4	100	50	0.81
LLDF with common features selected from 3 feature selection algorithm	64.3	83.3	50	0.79	57.1	66.7	50	0.62	71.4	100	50	o.81

S3

LC-MS/MS Protocol for Protein Identification from Blood Plasma Fraction

Sample Fractionation by Reverse-Phase HPLC

Blood plasma was precipitated with chilled acetonitrile (ACN) in a 1:1 ratio and centrifuged at 3,000 rpm for 5 minutes. The supernatant was collected and diluted 1:3 with HPLC-grade water. The diluted fraction was manually injected into a Thermo Fisher Dionex Ultimate 3000 HPLC system with a fluorescence detector. The Separation was performed on an Agilent Zorbax C-8 reverse-phase column (250 × 4.6 mm, 5 µm particle size, 300 Å pore size). Mobile phase A consisted of HPLC-grade water containing 0.1% trifluoroacetic acid (TFA), and mobile phase B consisted of acetonitrile containing 0.1% TFA. The resulting fraction was collected for downstream proteomic analysis.

Fraction Processing and Desalting

The collected HPLC fraction was dried in a vacuum centrifuge and reconstituted in 30 µL of 50 mM ammonium bicarbonate (pH 8.0). To remove the residual salts, the sample was desalted by spin-diafiltration using 10 kDa molecular weight cutoff filter.

Protein Digestion

Disulfide bonds were reduced with 10 mM dithiothreitol (DTT) at 37 °C for 30 minutes, followed by alkylation with 20 mM iodoacetamide at room temperature for 30 minutes in the dark. Trypsin was added at an enzyme-to-protein ratio of 1:50, and proteolytic digestion was carried out overnight at 37 °C. The resulting peptides were desalted using a C18 solid-phase extraction (SPE) column, concentrated, and reconstituted in 0.1% formic acid prior to LC-MS/MS analysis.

LC-MS/MS Analysis

Peptide separation was performed on an Agilent Infinity 1290 UHPLC system equipped with a reverse-phase C18 column. A 65-minute gradient was used at a flow rate suitable for the column dimensions. The UHPLC was coupled online to an Agilent 6545XT LC/Q-TOF mass spectrometer operating in positive ionization mode. Data were acquired using Auto MS/MS(Data-Dependent Acquisition, DDA), where full MS1 scans were followed by automatic selection and fragmentation of the most intense precursor ions. The instrument parameters were optimized for high-resolution accurate-mass analysis of tryptic peptides.

Data Processing and Protein Identification

Raw data were processed using PEAKS Studio V10.6 software. Protein identification was performed by searching against the Human Proteome database (UniProt). The key metrics included -10lgP scores, sequence coverage (%), label-free quantification (Area), number of unique peptides, and post-translational modifications (PTMs

Table S3

Protein(Unipot)	-10lgP	Coverage	Area(approx.)	No.of Unique peptides
ITPR1(Q14643)	53.01	1%	1.83×10^2	2
MSLN(Q13421)	22.87	1%	6.86×10^2	1
DSG2(Q14126)	28.16	1%	3.8×10^2	1
COL3A1(PQ2461)	20.64	1%	Low	1
CR2(P20023)	25.95	1%	Low	1

References

1. Leardi, Riccardo. "Genetic algorithms in feature selection." Genetic algorithms in molecular modeling. Academic Press, 1996. 67-86.
2. Farrés, Mireia,Stefan Tsakovski, and Romà Tauler. "Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation." Journal of Chemometrics 29, no. 10 (2015): 528-536.
3. Passos, Dário, and Puneet Mishra. "A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks." Chemometrics and Intelligent Laboratory Systems 223 (2022): 104520.
4. Khater, Omar, Ali Khater, Ashar Seif Al-Nasr, Samir Abozyd, Bassem Mortada, and Yasser M. Sabry. "Advancing near-infrared spectroscopy: A synergistic approach through Bayesian optimization and model stacking." Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 318 (2024): 124492.

5. Saito, Suguru, Yoshitoshi Hirao, Ali F. Quadery, Bo Xu, Amr Elguoshy, Hidehiko Fujinaka, Shohei Koma, Keiko Yamamoto, and Tadashi Yamamoto. "The optimized workflow for sample preparation in LC-MS/MS-based urine proteomics." Methods and Protocols 2, no. 2 (2019): 46.
6. Tsai, Tsung-Heng, Minkun Wang, and Habtom W. Ressom. "Preprocessing and analysis of LC-MS-based proteomic data." In Statistical analysis in proteomics, pp. 63-76. New York, NY: Springer New York, 2016.
7. Castillo, Sandra, Peddinti Gopalacharyulu, Laxman Yetukuri, and Matej Orešič. "Algorithms and tools for the preprocessing of LC–MS metabolomics data." Chemometrics and Intelligent Laboratory Systems 108, no. 1 (2011): 23-32.