DOI: 10.37990/medr.1718952 Med Records 2025;7(3):763-7

MEDICAL RECORDS-International Medical Journal

Research Article

Machine Learning-Driven Metabolomic Biomarker Discovery for PCOS: An Interpretable Approach Using Random Forest and SHAP

Seyma Yasar

İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye

Abstract

Aim: This study aimed to predict Polycystic Ovary Syndrome (PCOS) using follicular fluid metabolomic data and the Random Forest algorithm, and to interpret the contributions of the most influential metabolites using SHapley Additive exPlanations (SHAP) analysis.
Material and Method: An untargeted metabolomic dataset of follicular fluid from 35 PCOS patients and 37 age-matched controls was utilized. The dataset was partitioned into 70% training and 30% testing subsets using stratified sampling. A Random Forest algorithm was employed, with hyperparameter optimization performed using RandomizedSearchCV. Model performance was evaluated using accuracy, sensitivity, specificity, F1 score, balanced accuracy, and Brier score. SHAP analysis was then applied to interpret the model's predictions and identify key contributing metabolites.
Results: The Random Forest model achieved robust classification performance, with an accuracy of 0.86, sensitivity of 0.82, specificity of 0.91, F1 score of 0.86, balanced accuracy of 0.85, and a Brier score of 0.13. SHAP analysis identified L-Histidine, L-Glutamine, and L-Tyrosine as the top three most influential metabolites. Specifically, decreased levels of L-Histidine and L-Tyrosine, and elevated levels of L-Glutamine, were associated with an increased risk of PCOS.
Conclusion: Our findings demonstrate the potential of integrating machine learning with explainable AI to accurately predict PCOS based on metabolomic profiles. The identified metabolites, particularly alterations in amino acid metabolism, offer novel insights into the metabolic underpinnings of PCOS and highlight their promise as diagnostic biomarkers, paving the way for more precise and interpretable diagnostic strategies.

Keywords: Polycystic ovary syndrome, metabolomics, random forest, shapley additive explanations, biomarkers

INTRODUCTION

Polycystic ovary syndrome (PCOS) is one of the most common endocrine disorders in women, affecting approximately 10–15% of women of reproductive age. It is a complex, heterogeneous, and multifactorial condition (1,2). The key clinical features of PCOS include oligo/anovulation, hyperandrogenism, and polycystic ovarian morphology. Additionally, PCOS is frequently associated with metabolic abnormalities such as insulin resistance, dyslipidemia, type 2 diabetes, and increased cardiovascular risk (3,4). Due to its variable clinical presentation and underlying biological mechanisms, PCOS is increasingly recognized not as a single disease but as a spectrum of different phenotypes (5). Recent studies have suggested that PCOS results from an interplay between genetic predisposition, environmental influences, and lifestyle factors, involving hormonal, metabolic, and inflammatory pathways (6). Traditional biomarkers used for the diagnosis and classification of PCOS are often insufficient to capture this heterogeneity, complicating early diagnosis and the development of personalized therapeutic strategies. In this context, metabolomics- as a systems biology approach- has emerged as a powerful tool for identifying novel biomarkers and uncovering the metabolic subtypes of PCOS (7). Metabolomic studies have revealed significant alterations in energy metabolism, amino acid profiles, lipid metabolism, and oxidative stress pathways in individuals with PCOS (8,9).

However, analyzing and interpreting such high-dimensional and complex data requires advanced analytical approaches beyond conventional statistical methods. Machine learning (ML) algorithms, particularly ensemble

CITATION
Yasar S. Machine Learning-Driven Metabolomic Biomarker Discovery for PCOS: An Interpretable Approach Using Random Forest and SHAP. Med Records. 2025;7(3):763-7. DOI:1037990/medr.1718952

Received: 13.06.2025 **Accepted:** 22.07.2025 **Published:** 09.09.2025
Corresponding Author: Seyma Yasar, İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
E-mail: seyma.yasar@inonu.edu.tr

methods like Random Forest, have demonstrated strong performance in classification tasks and are increasingly applied to disease prediction using metabolomic data (10). The Random Forest algorithm is especially valuable in biomarker discovery due to its embedded feature selection and resistance to overfitting (11). Nevertheless, the interpretability of these complex models is of paramount importance in biomedical applications to ensure their clinical trustworthiness. In this regard, SHapley Additive exPlanations (SHAP) has become a widely used method for interpreting ML model predictions. SHAP allows for quantifying and visualizing the individual contribution of each feature to a given prediction, thereby enhancing model transparency and facilitating biological interpretation (12,13).

In this study, we aimed to predict PCOS using metabolomic data and the Random Forest algorithm, and to subsequently interpret the contribution of the most influential metabolites using SHAP analysis. This approach seeks to provide an accurate and interpretable diagnostic model while also offering novel insights into the metabolic underpinnings of PCOS.

MATERIAL AND METHOD

Dataset

We utilized an untargeted metabolomic dataset of follicular fluid from 35 PCOS patients and 37 age-matched controls (14). Ethical review and approval were not required for this study, in accordance with institutional requirements and national legislation, as the analyses were performed using an open-access dataset. This dataset provides comprehensive metabolite profiles and sample metadata under CC BY 4.0 license. In the PCOS group (n=35), the majority of participants (n=29, 82.9%) exhibited prolonged irregular menstrual cycles, categorized as "Irregular (longer)", while a smaller proportion (n=6, 17.1%) were diagnosed with oligomenorrhea.

Random Forest–Based Machine Learning Modeling Process

The dataset was partitioned into training (70%) and testing (30%) subsets using stratified sampling to preserve class distribution. Random Forest, an ensemble learning algorithm, was employed due to its robustness and ability to handle high-dimensional data with complex interactions. It constructs multiple decision trees on bootstrap samples of the training data, each considering a random subset of features, which helps reduce overfitting and improve generalization. Hyperparameter optimization for the Random Forest model was performed using RandomizedSearchCV with 10 iterations and 5-fold cross-validation (15,16). The optimized model demonstrated strong classification performance, evaluated by various metrics including accuracy, sensitivity (recall), specificity, F1 score, balanced accuracy, and Brier score. To enhance model interpretability and elucidate the decision-making process, SHAP values were calculated. SHAP is a model-agnostic interpretability technique based on cooperative game theory that quantifies the contribution of each feature to individual predictions. This approach allowed the identification and visualization of the most influential metabolites involved in the prediction of PCOS, providing valuable biological insights and supporting the clinical applicability of the model (17).

Statistical Analysis

Data were summarized as mean±standard deviation or median (interquartile range; IQR). Normality was assessed separately for each group using the Shapiro-Wilk test. For variables meeting the normality assumption, group differences were evaluated using the independent two-sample t-test, while variables not meeting the assumption were analyzed with the Mann-Whitney U test. A p-value of less than 0.05 was considered statistically significant. Data analysis was performed using IBM SPSS Statistics for Windows, Version 26.0 (18) and the Python programming language (19).

RESULTS

The dataset used in this research consists of a total of 72 patients, 35 PCOS (48.6%) and 37 (51.4%). The mean age of all patients is 30.2±3.9 years and the median age is 29 (23-41). Table 1 presents the baseline demographic features of individuals in the PCOS and control groups.

Table 1. The comparison of demographic characteristics between PCOS and control groups

	Group				p-value
	Control (n=37)		PCOS (n=35)		
	Mean±SD	Median (IQR)	Mean±SD	Median (IQR)	
Age (years)	31.11±4.18	31 (5.5)	29.31±3.38	29 (3)	**0.040***
Height (m)	1.62±0.04	1.63 (0.05)	1.61±0.05	1.6 (0.07)	0.213**
Weight (kg)	57.71±7.83	55 (13.5)	58.77±8.61	60 (10)	0.619**
BMI (kg/m²)	22.07±2.84	21.99 (4.8)	22.73±2.99	22.15 (4.4)	0.434*
Serum testosterone (ng/ml)	0.41±0.1	0.45 (0.19)	0.54±0.21	0.5 (0.32)	**0.007***

SD: standard deviation, IQR: interquartile range, PCOS: polycystic ovary syndrome, BMI: body mass index; *: Mann-Whitney U; **: Independent sample t-test

The Random Forest algorithm applied to the metabolites obtained from the metabolomic analyses yielded the following performance metrics: accuracy=0.86, sensitivity (recall)=0.82, specificity=0.91, F1 score=0.86, balanced accuracy=0.85, and Brier score=0.13. These results indicate that the model achieved a high level of predictive performance, with a good balance between sensitivity and specificity, suggesting its potential utility in distinguishing individuals with PCOS from controls based on metabolomic profiles. Following SHAP analysis to enhance model interpretability, the top 3 most influential metabolites contributing to the classification were identified as L-Histidine, L-Glutamine, and L-Tyrosine, highlighting the potential metabolic signatures associated with PCOS (Figure 1). Table 2 presents the distribution and statistical differences of the top 3 key metabolites identified by the model across the groups. The distribution and directionality of feature contributions to the model's predictions were further visualized using a SHAP beeswarm plot (Figure 2), which illustrates the impact of each metabolite on individual predictions and highlights their relative importance across the dataset.

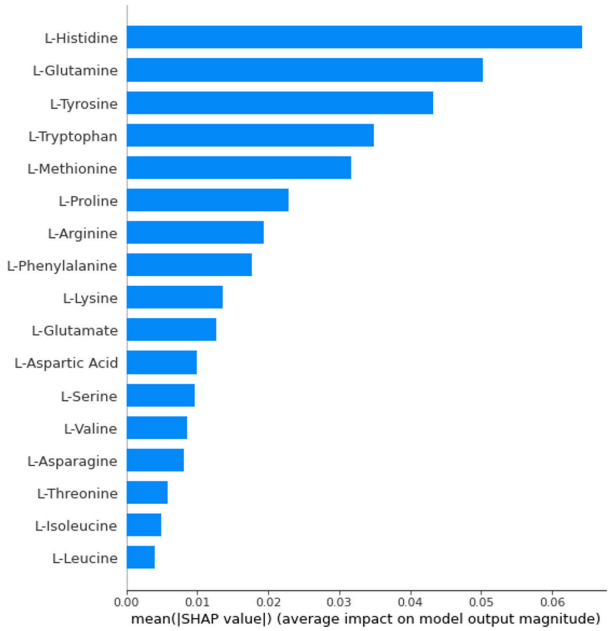

Figure 1. The metabolites contributing to the classification of PCOS versus control individuals, as identified through SHAP analysis

Table 2. Comparison of the distributions and statistical differences of the top 10 most important metabolites across the relevant groups

	Group				p-value
	Control		PCOS		
	Mean±SD	Median (Min-Max)	Mean±SD	Median (Min-Max)	
L-Histidine	276826.07±61826.56	259220.76 (161017.47-432697.39)	229731.58±32052.18	230248.32 (167998.47-314654.97)	**<0.001***
L-Glutamine	143904.12±22185.97	145987.85 (100208-179231.45)	171716.71±33932.91	166242 (127857-287643)	**<0.001***
L-Tyrosine	18593.45±4130.27	17175.95 (13129.78-29317.8)	15693.43±2754.86	15805.95 (11079.54-21710.02)	**0.005***

SD: standart deviation, IQR: interquartile range, PCOS: polycystic ovary syndrome; *: Mann-Whitney U

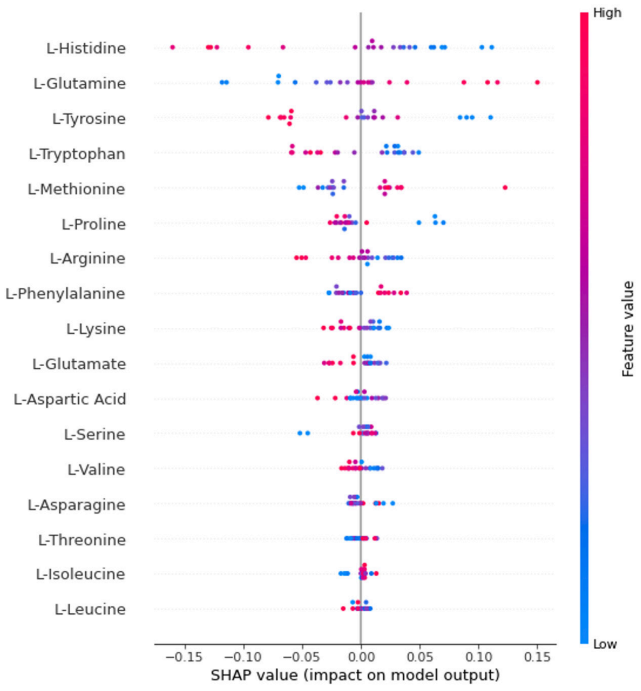

Figure 2. SHAP beeswarm plot illustrating the distribution and directionality of feature contributions to the model's predictions. Each point represents a single sample, showing the impact of individual metabolites on prediction outcomes and highlighting their relative importance across the dataset

DISCUSSION

PCOS is a complex endocrine disorder that is commonly observed in women of reproductive age and is characterized by heterogeneous clinical manifestations. Although the exact pathophysiology of PCOS has not been fully elucidated, insulin resistance, hormonal imbalances, and chronic inflammation are known to play central roles in its development. In recent years, multi-omics approaches and high-dimensional biomarker data have gained increasing importance in efforts to better understand the biological basis of PCOS. In this context, machine learning and artificial intelligence techniques have emerged as powerful tools for identifying meaningful patterns and potential biomarkers from large-scale metabolomic datasets (20). Supporting this perspective, Nsugbe (2023) proposed an AI-driven clinical decision support system for the early diagnosis and staging of PCOS. The study addressed common limitations in the field, such as class imbalance in training data, by applying synthetic oversampling (SMOTE) to the widely used Kaggle PCOS dataset. Ten machine learning models were evaluated, and nonlinear support vector machine (SVM) variants-particularly Cubic and Medium Gaussian SVMs-demonstrated the highest classification performance. Notably, the study extended beyond

binary classification to include a probabilistic inference system capable of distinguishing early and advanced stages of the condition. These findings highlight the potential of AI not only for accurate PCOS diagnosis but also for informing proactive, stage-specific clinical management strategies (21). Supporting this, Elmannai et al. (2023) developed a comprehensive PCOS prediction framework combining optimized feature selection and explainable AI (XAI) methodologies. Their study employed multiple machine learning algorithms, including logistic regression, random forest, SVM, and XGBoost, optimized via Bayesian search, and addressed class imbalance using a hybrid SMOTE-ENN strategy. Notably, the stacking ensemble model incorporating Recursive Feature Elimination (RFE) achieved 100% accuracy in the 80:20 data split and maintained high performance across metrics (ACC=98.87%, PRE=98.00%, REC=98.87%, F1=98.89%) in 70:30 splitting. In addition to predictive success, the study emphasized model transparency by integrating both global and local explanation techniques, demonstrating the importance of features such as follicle number and hormone levels. These findings affirm the growing utility of AI-based systems not only for accurate PCOS detection but also for interpretable, reliable, and individualized risk assessment (21). In this study, a Random Forest classification model based on metabolomic data was developed to identify potential biomarker metabolites for the diagnosis of PCOS, and the model's decision mechanisms were interpreted using SHAP analysis. The Random Forest model demonstrated a robust performance in distinguishing PCOS from control samples based on metabolomic data, with an accuracy of 0.86, sensitivity of 0.82, specificity of 0.91, F1 score of 0.86, balanced accuracy of 0.85, and a Brier score of 0.13. SHAP analysis was employed to interpret the model's decision-making process and to identify key metabolites contributing to the classification. Among all features, L-Histidine, L-Glutamine, and L-Tyrosine emerged as the top three most influential metabolites. Notably, decreased levels of L-Histidine and L-Tyrosine were associated with an increased risk of PCOS, while elevated levels of L-Glutamine were similarly linked to a higher PCOS risk. These findings suggest that specific alterations in amino acid metabolism may play a crucial role in the pathophysiology of PCOS and highlight their potential as diagnostic biomarkers. SHAP analysis identified L-Histidine, L-Glutamine, and L-Tyrosine as the top three metabolites influencing PCOS classification. Decreased plasma L-Histidine levels in PCOS have been consistently observed in multiple studies, likely reflecting its role as an antioxidant consumed under oxidative stress, with histidine showing high diagnostic power (AUC= 0.90) in adipose tissue and plasma samples. Elevated L-Glutamine levels have also been reported in both fasting adolescent PCOS patients (p=0.04) and DHEA-induced PCOS rat models; in the latter, supplementation modulated inflammation, oxidative stress, insulin resistance (HOMA-IR) and LH levels (22-24). In our model, increased L-Glutamine was associated with higher PCOS risk, aligning with its implication in metabolic dysregulation. L-Tyrosine typically shows increased concentrations in PCOS follicular fluid and exhibits positive correlations with androgen levels. Interestingly, our model indicated that reduced L-Tyrosine corresponded to elevated PCOS risk- a finding that diverges from existing literature and might be explained by differences in sample type, phenotypic subgroups, or the model's interpretative framework. Taken together, these results underscore the pivotal role of amino acid metabolism in PCOS pathophysiology and support further investigation of these metabolites as candidate diagnostic biomarkers. Overall, this study demonstrates the potential of integrating machine learning with explainable AI to uncover metabolite-based biomarkers, offering new insights into the metabolic underpinnings of PCOS and paving the way for more precise, interpretable, and personalized diagnostic strategies.

CONCLUSION

The findings of this study demonstrated that the integration of machine learning with explainable AI can provide accurate and interpretable diagnostic models for PCOS. The Random Forest model showed high classification performance, while SHAP analysis revealed that alterations in amino acid metabolism- particularly involving L-Histidine, L-Glutamine, and L-Tyrosine- play a significant role in distinguishing PCOS from controls. These findings highlight the potential of metabolomic biomarkers for improving diagnostic accuracy and deepening our understanding of the metabolic mechanisms underlying PCOS. Further validation in larger, independent cohorts is warranted to confirm the clinical utility of these metabolites and support their translation into personalized diagnostic strategies.

Financial disclosures: *The authors declared that this study has received no financial support.*

Conflict of interest: *The authors have no conflicts of interest to declare.*

Ethical approval: *As the research utilized only publicly available open-access data, ethical approval was not required under institutional and national guidelines.*

REFERENCES

1. Teede HJ, Tay CT, Laven JJ, et al. Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Eur J Endocrinol. 2023;189:G43-64.
2. Lizneva D, Suturina L, Walker W, et al. Criteria, prevalence, and phenotypes of polycystic ovary syndrome. Fertil Steril. 2016;106:6-15.
3. Goodarzi MO, Korenman SG. The importance of insulin resistance in polycystic ovary syndrome. Fertil Steril. 2003;80:255-8.
4. Dumesic DA, Abbott DH, Chazenbalk GD. An evolutionary model for the ancient origins of polycystic ovary syndrome. J Clin Med. 2023;12:6120.

5. Chang K-J, Chen J-H, Chen K-H. The pathophysiological mechanism and clinical treatment of polycystic ovary syndrome: a molecular and cellular review of the literature. Int J Mol Sci. 2024;25:9037.

6. Sharma I, Dhawan C, Arora P, et al. Role of environmental factors in PCOS development and progression. Herbal Medicine Applications for Polycystic Ovarian Syndrome: CRC Press. 2023;281-300.

7. Xuan Y, Hong X, Zhou X, et al. The vaginal metabolomics profile with features of polycystic ovary syndrome: a pilot investigation in China. PeerJ. 2024;12:e18194.

8. Liu R, Bai S, Zheng S, et al. Identification of the metabolomics signature of human follicular fluid from PCOS women with insulin resistance. Dis Markers. 2022;2022:6877541.

9. Alesi S, Ghelani D, Mousa A. Metabolomic biomarkers in polycystic ovary syndrome: a review of the evidence. Semin Reprod Med. 2021;39:102-10.

10. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, et al. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. Front Microbiol. 2021;12:634511.

11. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform. 2019;20:492-503.

12. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56-67.

13. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017;30.

14. Yuan Y. DataSet for PCOS. Mendeley Data. 2023;V1. doi: 10.17632/mh94mxn3nh.1.

15. Vishnu M, Rupak VV, Vedhapriyaa S, et al., Recurrent gastric cancer prediction using randomized search cv optimizer. 2023 International Conference on Computer Communication and Informatics (ICCCI), 23-25 Jan. 2023. Coimbatore, India, 1-5.

16. Xie N-N, Wang F-F, Zhou J, et al. Establishment and analysis of a combined diagnostic model of polycystic ovary syndrome with random forest and artificial neural network. Biomed Res Int. 2020;2020:2613091.

17. Van den Broeck G, Lykov A, Schleich M, Suciu D. On the tractability of SHAP explanations. Journal of Artificial Intelligence Research. 2022;74:851-86.

18. IBM Corp. SPSS Statistics for Windows. V. 26.0. IBM Corp Armonk, NY; 2019.

19. Srinath K. Python–the fastest growing programming language. International Research Journal of Engineering and Technology. 2017;4:354-7.

20. Verma P, Maan P, Gautam R, Arora T. Unveiling the role of artificial intelligence (AI) in polycystic ovary syndrome (PCOS) diagnosis: a comprehensive review. Reprod Sci. 2024;31:2901-15.

21. Nsugbe E. An artificial intelligence-based decision support system for early diagnosis of polycystic ovaries syndrome. Healthcare Analytics. 2023;3.100164.

22. Di F, Gao D, Yao L, et al. Differences in metabonomic profiles of abdominal subcutaneous adipose tissue in women with polycystic ovary syndrome. Front Endocrinol (Lausanne). 2023;14:1077604.

23. Wu G, Hu X, Ding J, Yang J. The effect of glutamine on Dehydroepiandrosterone-induced polycystic ovary syndrome rats. J Ovarian Res. 2020;13:57.

24. Cree-Green M, Carreau A-M, Rahat H, et al. Amino acid and fatty acid metabolomic profile during fasting and hyperinsulinemia in girls with polycystic ovarian syndrome. Am J Physiol Endocrinol Metab. 2019;316:E707-18.