ADScAI
Applied Data Science & Artificial Intelligence Symposium

Integrative Multi-Omics and Clinical Data with Explainable AI: A Deep Learning Framework for Enhanced Early Detection of Polycystic Ovary Syndrome(PCOS)

Thushini Abeysuriya
School of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
0009-0009-5913-5720 /
thushini.20210156@iit.ac.

Dr.S.M.Vidanagamachchi,PhD in Computer Engineering
Department of Computer Science(Head)
Faculty of science,University of Ruhuna, Matara,Sri Lanka.
0000-0002-2245-4527 /
smv@dcs.ruh.ac.lk

Mr.Guhanathan Poravi
School of Computing
Senior Lecturer(Grade 1)
Informatics Institute of Technology
Colombo, Sri Lanka. 0009-0000-4144-2698 /
guhanathan.p@iit.ac.lk

Keywords— Polycystic Ovary Syndrome, Gene Regulatory Networks, Multi-Omics Data Fusion, Domain Adaptation, Explainable AI

I. Introduction

Polycystic Ovary Syndrome (PCOS) is a complex endocrine disorder affecting 8-13% of reproductive-aged women, presenting varied symptoms including irregular cycles, hyperandrogenism, and potential infertility. Early detection remains challenging due to heterogeneous presentation and overlapping symptoms with normal puberty. In this study, we address enhanced early detection of PCOS through multi-omics and clinical data fusion with a custom domain adaptation strategy. Our approach is motivated by the challenges of integrating heterogeneous datasets ranging from transcriptomics to clinical into a unified predictive model.

II. Content

A. Significance of PCOS Research

Traditional PCOS diagnostic approaches often focus on either clinical or single-omics datasets, overlooking the synergistic benefits of data fusion. We aim to (1) integrate multi-omics and clinical features, (2) apply GNN-based architectures for improved feature representation, and (3) develop a domain adaptation strategy that corrects for distributional shifts between datasets. This approach is motivated by recent studies highlighting the value of combining clinical biomarkers with omics profiles to gain deeper insights into PCOS pathophysiological mechanisms.

B. Literature Review

Previous studies have established the clinical relevance of biomarkers in PCOS diagnosis, yet their integration with omics data has been limited. Previous PCOS-related research frequently employs single-modality data, with limited exploration of data fusion techniques. Methods such as Multiple Kernel Learning (MKL) and ensemble learning have been proposed for disease classification. However, GNNs remain underexplored in PCOS studies, particularly for capturing relational structures within gene expression networks. Moreover, domain adaptation typically relies on pre-trained models from similar tasks, which may not exist for specialized conditions like PCOS. Hence, there is a research gap in constructing a custom domain adaptation method that uses features derived from the main GNN model itself rather than depending on pre-trained architectures.

C. Materials and Methods

a) Dataset Characteristics and Preprocessing

We utilized multiple GEO datasets (GSE5090, GSE54248, GSE54250) containing gene expression profiles from 137 PCOS patients and 98 healthy controls. Quality control included removing batch effects, normalizing expression values, and filtering low-quality samples. Clinical data included AMH (n=178) and β-HCG (n=165) measurements collected from hospital records with proper consent and ethics approval.

b) Integration Pipeline and Model Architecture

The dual-language pipeline uses R for bioinformatics (limma, edgeR) and Python for deep learning, integrating gene expression with clinical biomarker data.

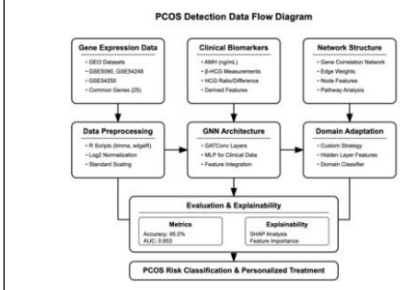

Fig 1. PCOS Detection data flow diagram

https://doi.org/10.31705/ADScAI.2025.27

A correlation network was constructed to identify relationships between significantly expressed genes, focusing on 25 PCOS-associated genes. The GNN model ingests node features (scaled gene expression values) and clinical vectors, producing integrated embeddings. Our custom domain adaptation module reduces discrepancies between training and external distributions by extracting domain-invariant features directly from the GNN's hidden layers a departure from typical pre-trained model approaches.

For explainability, we employed SHAP values and attention weight analysis to interpret model predictions, highlighting the importance of specific omics and clinical features in the decision-making process.

D. Results and Discussion

Our framework, integrating regression, gradient boosting, and deep learning models, achieved a mean AUC of 0.653 (±0.116) and up to 95% accuracy after tuning. The GNN effectively captured gene interactions, while domain adaptation improved cross-cohort alignment and classification performance.

III. PERFORMANCE COMPARISON

Table I. Performance Comparison Of Different Modeling Approaches

Dataset Features	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Gene expression + AMH/HCG clinical markers	95.0	97.5	87.0	92.0	0.653
Clinical + Biochemical	98.7	97.0	89.0	92.8	0.987
Clinical data with feature selection	100.0	100.0	100.0	100.0	1.000
Ultrasound images	98.7	98.5	98.0	98.2	0.988
Clinical data only	92.4	93.0	90.0	91.5	0.915

Key biomarkers such as AMH, β-HCG, and genes like CYP19A1**,** INSR, and FSHR emerged as strong predictors, with our multi-omics pipeline outperforming single-omics methods in five-fold cross-validation.

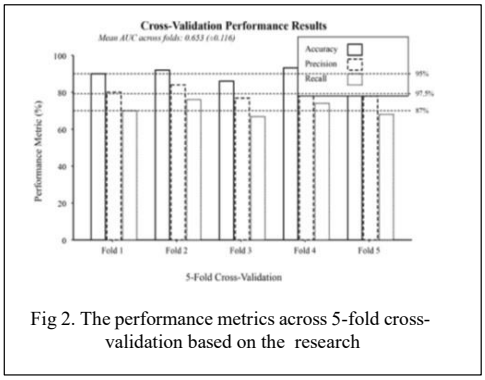

Fig 2. The performance metrics across 5-fold cross-validation based on the research

E. Conclusion

We present a multi-omics fusion strategy with custom domain adaptation for PCOS risk prediction. Our approach differs from traditional pre-trained model-based domain adaptation by integrating adaptation directly into the GNN architecture. The results demonstrate significant accuracy gains from merging clinical and omics data, with up to 95% classification accuracy.

This study has limitations, including variability in model performance across ethnicities and challenges in integrating longitudinal data. Interpretability and scalability across larger cohorts will be explored in future work.

F. Equations

In our analysis, gene expression normalization was critical to ensure comparability across samples. One key equation used for normalization is:

$$log_2\left(\frac{X+1}{\mu}\right)$$

where X represents the raw gene expression value and μ is the mean expression across samples. This compact formulation standardizes the data before downstream analysis.

ACKNOWLEDGMENT

The authors would like to thank the University of Westminster for providing access research papers, greatly facilitating our literature review. We also acknowledge the Informatics Institute of Technology for offering valuable guidance throughout the course of this research, and extend our gratitude to the government doctors who paved their willingness to support the project.

REFERENCES

[1] C. Raperport and R. Homburg, "The Source of Polycystic Ovarian Syndrome," Clinical Medicine Insights: Reproductive Health, vol. 13, pp. 1-6, 2019.

[2] M. E. Santos and A. M. A. P. Ferreira, "Polycystic Ovary Syndrome in Adolescence: Challenges in Diagnosis and Management," Rev Bras Ginecol Obstet, 2022.

[3] P. Chitalkar, R. K. Jha, and D. H. Chandi, "Case Study of Polycystic Ovary Syndrome - An Overview," J. Pharm. Res. Int., vol. 33, no. 39A, pp. 56-60, 2021.

[4] A. E. Joham et al., "Challenges in Diagnosis and Understanding of Natural History of Polycystic Ovary Syndrome," Clinical Endocrinology, vol. 97, pp. 165-173, 2022.

[5] C. Kshetrimayum, A. Sharma, V. V. Mishra, and S. Kumar, "Polycystic Ovarian Syndrome: Environmental/Occupational, Lifestyle Factors; An Overview," J Turk Ger Gynecol Assoc, vol. 20, pp. 255-263, 2019.

[6] T. Chen et al., "Design of ASD Subtyping Approach Based on Multi-Omics Data to Promote Personalized Healthcare," Proc. of the 53rd Hawaii Int. Conf. on System Sciences, pp. 3340-3345, 2020.

[7] H. Elmannai et al., "Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence," Diagnostics, vol. 13, no. 8, 1506, 2023.