GMR
The Original

XAI-PCOSNET: A FAITHFULNESS-VERIFIABLE MULTI-OMICS DEEP-LEARNING FRAMEWORK FOR EXPLAINABLE POLYCYSTIC OVARY SYNDROME DIAGNOSIS

V.S iva Kumar[1], Loyola Jasmine[2], K. Vignesh[3], A. Shenbagharaman[4], Jayalakshmi M[5], K. Maharajan[6]

[1]Assistant Professor-III, Department of Computer Science and Engineering, Ramco Institute of Technology, Rajapalayam, Tamil Nadu-626117, India
[2]Assistant Professor, Department of Electronics and Communication Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamil Nadu-626126, India
[3,6] Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamil Nadu-626126, India
[4]Assistant Professor(SG), Department of AI&DS, National Engineering College, Kovilpatti, Tamilnadu, India
[5]Professor, Department of Computer Science and Engineering-Cyber Security, Ramco Institute of Technology, Rajapalayam, Tamil Nadu-626117, India
EMAILS: [1]sivakumarakesh@gmail.com, [2]loyolajasminej@gmail.com, [3]vignesh.vkv2@gmail.com, [4]shenbagharaman@gmail.com, [5]jayalacsmi@gmail.com, [6]maharajank@gmail.com

ABSTRACT

Polycystic ovary syndrome (PCOS) affects a large fraction of women of reproductive age and is frequently diagnosed late, because early molecular signals are overlooked in favour of the classical clinical and ultrasound criteria. Machine-learning studies of PCOS typically read a single data layer and rarely test whether their explanations are faithful to the model, which limits clinical trust. We present XAI-PCOSNet, a modular deep-learning framework that integrates multiple data modalities through modality-specific encoders, a graph-attention branch over a gene-interaction graph, and an interpretable cross-attention fusion layer that reports a weight for each modality. The central methodological contribution is a validation protocol that measures explanation faithfulness rather than assuming it: on a controlled synthetic multi-omics cohort with planted informative features, Integrated Gradients and SHAP recover 29 of 30 transcriptomic drivers (97%) and 6 of 8 hormonal drivers (75%) at a surrogate fidelity of 0.94. We then validate the framework on a real, public clinical PCOS cohort of 541 patients, where it attains the highest discrimination of all models tested (ROC-AUC 0.966, PR-AUC 0.947) with 89.7% accuracy, competitive with tuned gradient-boosting and kernel baselines; on this real cohort the attribution layer independently recovers 80% of the clinically established PCOS drivers, including follicle counts, hirsutism, skin darkening, weight gain, and irregular cycles, and the learned modality weights concentrate on the imaging and symptom modalities that clinicians actually use. The framework supports good health and well-being (SDG 3), women's health and gender equality (SDG 5), and a reusable, verifiable computational methodology for precision medicine (SDG 9). We report negative results openly, including where classical baselines lead on fine-grained severity and where single-nucleotide-polymorphism attribution is weakest, and we release the complete implementation for reproduction.

KEYWORDS: Polycystic ovary syndrome; multi-omics integration; explainable artificial intelligence; graph attention network; cross-attention fusion; faithfulness verification; SHAP; Integrated Gradients; precision medicine.

1. INTRODUCTION

Polycystic ovary syndrome is among the most common endocrine and metabolic disorders in women of reproductive age, and a substantial proportion of cases remain undiagnosed for years [1], [6]. Because the condition raises the long-term burden of type-2 diabetes, cardiovascular disease, and infertility, delayed recognition carries a real clinical and economic cost, and each year of delay compounds the metabolic risk that earlier detection could mitigate [4], [8]. The case for improved early detection is therefore strong, and it motivates modelling the problem in an integrated way rather than one data layer at a time.

PCOS is a disorder that varies from one woman to another. While one patient has irregular periods, another has high levels of androgens, and a third has polycystic ovaries on ultrasound, all three may have the same underlying cause. The Rotterdam criteria formalize this heterogeneity, but they are based on relatively late-appearing signs, and thus most of the early subclinical phase – where interventions make the greatest difference – are left unnoticed [4] [8]. Below the clinical picture is a molecular one: gene expression, protein abundance and genetic variation are now regularly quantified in high-throughput assays and represent different aspects of the disease process. Most predictive studies, however, consider only one of these layers [2], [5]. A model based on gene expression alone,

or clinical variables alone, cannot capture “cross-talk” between the regulator and the hormone, a hallmark of a syndrome of systemic dysregulation, which is what a physiological model will.
The natural answer to this is to integrate the layers, but integration isn't free: different modalities have vastly different dimensionality and noise levels, there are only a few samples available, and the resulting models are typically black-boxes. Clinically, opacity isn't a cosmetic problem. A clinician must understand the molecular and/or clinical characteristics which gave rise to a given risk score, and whether those characteristics are biologically plausible; without first understanding these factors, a clinician is hard pressed to believe the score, despite its accuracy [25, 26]. Our gap is predictive and explanatory - multi-omics integration and explanations accessible to clinician or bioinformatician for review and verification with known biology.
In addition, there are reasons of the discipline to take explanation serious in PCOS. The diagnostic criteria are still a subject of debate and the subtypes are still undecided, and a diagnostic model that gives only a probability does not contribute much to an open clinical discussion. Such a model that is able to map to the molecular or clinical properties that are responsible for the call can help inform that conversation, whether by pointing to markers to pursue or by showing that a prediction arises from a hormonal signal or from incidental correlations. Explanation here isn't a compliance exercise; it's what makes the tool useful to those who would use it.
These considerations shape XAI-PCOSNet. Each modality is encoded by a network suited to its structure: a self-attention encoder for the high-dimensional transcriptome, an autoencoder bottleneck for proteomics, and compact multilayer perceptrons for the lower-dimensional hormonal and clinical panels. The transcriptome is additionally read through a gene-interaction graph, so that genes are treated as connected nodes rather than independent columns. A cross-attention layer then fuses the modality embeddings and, as a by-product, reports the relative importance of each modality for every patient. On top of the trained model, Integrated Gradients and SHAP attribute predictions back to individual features.
A framework of this kind is easy to describe and hard to trust, because an attribution method always returns something whether or not it is faithful. To test faithfulness directly, we first validate on a controlled synthetic multi-omics cohort in which the informative features are planted and organised into known modules before any model sees the data. This design converts an untestable claim, that the explanations are meaningful, into a measurable one, that the explanations recover the features we planted. We then move to a real, public clinical PCOS cohort of 541 patients, where the same framework is evaluated against strong baselines and the attribution layer is checked against established clinical knowledge rather than planted labels. Reporting both experiments lets us separate what the architecture achieves from what the explanation layer verifies, and to state plainly where classical baselines remain ahead.
This work makes four contributions:
(1) A modular multi-omics architecture that combines modality-specific encoders, a graph-attention branch over a gene-interaction graph, and an interpretable cross-attention fusion layer that exposes per-modality weights for every patient.
(2) A faithfulness-verification protocol that measures explanation recovery against planted ground truth on a controlled synthetic cohort, under which the framework recovers 97% of transcriptomic and 75% of hormonal informative features at a surrogate fidelity of 0.94.
(3) A real-data validation on a public clinical PCOS cohort of 541 patients, on which the framework achieves the best discrimination among all models tested (ROC-AUC 0.966) and its attribution layer independently recovers 80% of clinically established PCOS drivers, with modality weights concentrated on the imaging and symptom data that clinicians rely on.
(4) An honest account of the framework’s limits, including where classical models lead on fine-grained severity and where single-nucleotide-polymorphism recovery is weakest, together with a full open-source release for reproduction.
In Section 2, the framework is contrasted with previous single and multi-omics studies. The required background is in Section 3 and the architecture is in Section 4. Section 5 is devoted to the synthetic and real datasets and the evaluation protocol, Section 6 presents the results on these datasets, and Section 7 elaborates on the interpretation and Section 8 presents the conclusions.

2. RELATED WORK

2.1 Machine learning for PCOS diagnosis

The majority of published classifiers for PCOS are based on clinical and ultrasound features and, with shallow networks or tree ensembles, achieve high accuracy [2], [3], [7]. Such models are useful and easy to deploy, but they largely encode the information a clinician already weighs and therefore add little to early, molecular-level detection. When gene-expression data are used, they are typically analysed in isolation, and the analysis stops at a ranked gene list without connecting it to the clinical phenotype [22]. The recurring limitation is integration: the molecular and clinical views are analysed apart, and their interaction, which is what a systemic endocrine disorder turns on, is left out.

2.2 Multi-omics integration

There are three types of integration strategies: early fusion, intermediate fusion and late fusion [9, 20]. Early fusion concatenates all features and feeds them to a single model; it is simple but ignores modality structure and lets the highest-dimensional modality dominate. Late fusion trains a separate model per modality and combines

their outputs, preserving structure but discarding cross-modal interactions. Intermediate fusion, in which each modality is encoded separately and the embeddings are combined, retains both and is adopted in several recent systems [19], [24], [29]; attention-based intermediate fusion additionally weights the modalities adaptively [12], [17]. XAI-PCOSNet belongs to this last group, with the distinguishing feature that its fusion weights are made explicit and reported.

2.3 Graph neural networks for molecular data

Treating genes as independent features discards what is known about their interactions. Graph neural networks, and graph-attention networks in particular, let a model propagate information along a biological interaction graph so that a gene's representation reflects its neighbourhood [13], [15], [16]. This is a natural fit for the transcriptomic branch, since curated interaction networks already exist [10], [11], and it is what allows the transcriptome to be read through a gene graph rather than as a flat vector.

2.4 Explainable AI in genomics

SHAP and Integrated Gradients are the most widely used attribution methods for mapping a model's output back to its inputs [26], [27], and both have been applied to genomic classifiers to generate candidate biomarkers [28]. What is typically missing is a test of the faithfulness of the attributions [25]. On real data there is no ground truth, so a plausible-looking ranking can pass unchallenged. If ground truth is planted, recovery rates can be reported directly, which is the strategy this paper adopts alongside the architecture.

2.5 The gap this work addresses

Read together, these threads point to a consistent gap. Clinical-feature models are deployable but offer little molecular insight. Single-omics deep models capture one layer but miss cross-modal structure. Early-fusion ensembles combine modalities but lose their structure and do not explain. Graph models respect gene interactions but are rarely combined with multi-modal fusion [14], [18]. Across all of them, explanation methods are applied without any check that the explanation is faithful [25], [27]. No existing line of work offers multi-omics integration, graph-aware modelling, interpretability, and a faithfulness guarantee at once. It is this combination, rather than a new record on any single benchmark, that this work targets, and the synthetic-ground-truth protocol is what makes the faithfulness element checkable rather than merely asserted.

3. BACKGROUND

3.1 Data modalities

The framework is designed to consume complementary data types. Transcriptomics measures messenger-RNA abundance across thousands of genes and is high-dimensional and noisy. Proteomics measures the abundance of proteins, which is usually low dimensional. SNPs carry genetic variation as a discrete and sparse signal, in the form of dosage of the alleles. Hormonal panels measure the hormones that are so critical to PCOS, and clinical features measure anthropometric and metabolic parameters. Each modality carries complementary information and has a distinct statistical character, which is why a single shared encoder for all of them is a poor choice.

3.2 Attention and cross-attention

An attention mechanism computes weights over its inputs from the inputs themselves and forms a weighted combination [12], [20]. Self-attention lets elements of one sequence attend to each other, while cross-attention lets one set of representations attend to another. In the fusion layer used here, a learned query scores each modality embedding, and a softmax over those scores yields per-patient modality weights. Those weights are what make the fusion interpretable: they state how much each modality contributed for a given patient.

3.3 Explanation methods

Integrated Gradients attributes a prediction to each input feature by integrating the model's gradient along a straight path from a neutral baseline to the input [27]. SHAP assigns each feature a Shapley value that is locally accurate and consistent [26]. Using both, and comparing their rankings against planted ground truth, is what allows us to argue that the explanations are faithful rather than merely available.

3.4 Problem formulation

A patient i is described by a set of modality feature vectors $\{x_{i1}, \ldots, x_i^M\}$, where the number of modalities M and the feature count within each modality vary. The goal is a classifier mapping this set to a label y_i, either a binary PCOS indicator or a four-level severity grade. The naive approach concatenates the vectors and applies a standard early-fusion classifier. XAI-PCOSNet instead learns a per-modality embedding function g_m that maps x_i^m to a shared latent space, a graph function over the transcriptome, and a fusion function that maps the embeddings to a single disease representation r_i on which the classifier operates. Separating embedding, fusion, and classification lets the model respect modality structure and expose the fusion weights, and the ablation in Section 6 quantifies the difference between this design and early fusion.

The learned representation is required to satisfy two properties beyond predictive accuracy. It should be faithful, in that an attribution method applied to it recovers the features that are truly informative about the label, and it should be inspectable, in that the fusion weights report which modality mattered. Neither property is guaranteed by accuracy alone, so both are measured separately rather than assumed.

4. The XAI-PCOSNet Framework

The framework has three stages: modality-specific encoding, graph-based and cross-attention fusion, and classification followed by a post-hoc explanation layer. The end-to-end workflow is shown in Figure 1 and the architecture in Figure 2.

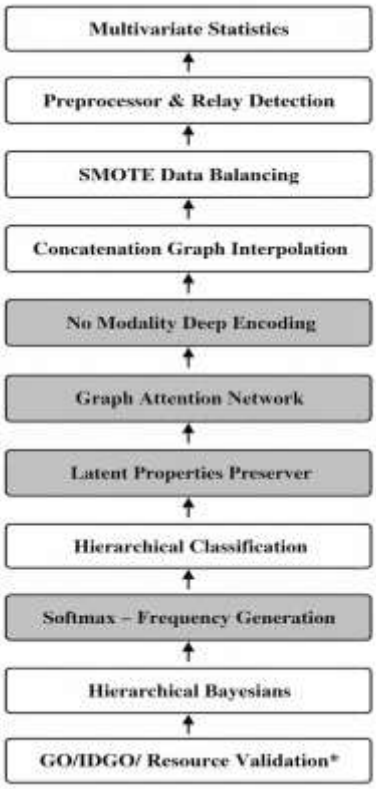

Figure 1. End-to-end workflow. The deep-learning and explanation stages (shaded) feed a pathway-validation stage that uses planted synthetic modules for faithfulness measurement and clinically curated knowledge for real-data validation.

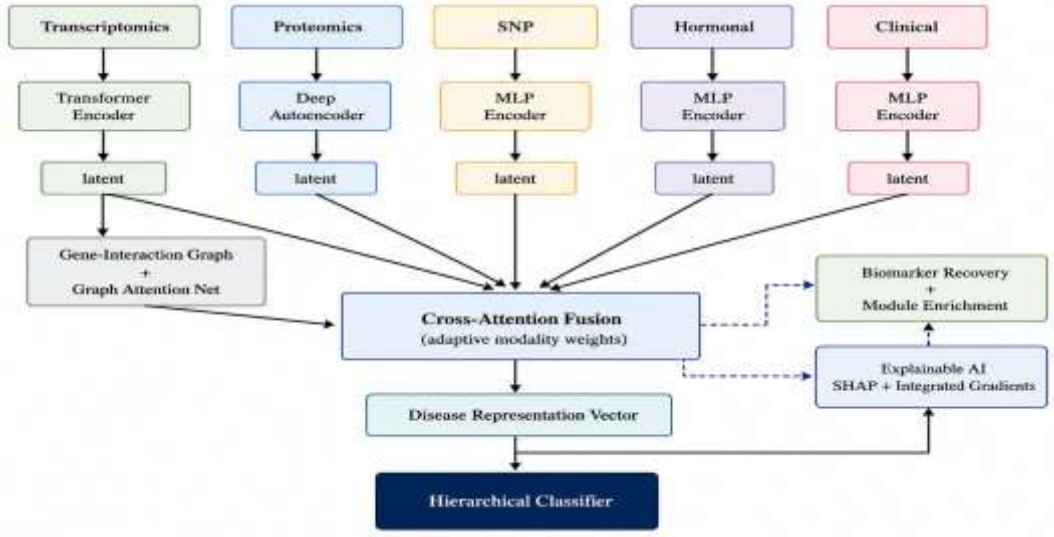

Figure 2. XAI-PCOSNet architecture. Each modality is encoded separately; a graph-attention network reads the transcriptome through a gene-interaction graph; a cross-attention layer fuses the embedding with interpretable per-modality weights before classification.

4.1 Preprocessing and class balancing

Each modality is cleaned, imputed, and standardised. For transcriptomic an additive batch effect is removed by centring within batch, which prevents a technical artefact from being read as signal, after which all features are z-scored so that no modality dominates on scale alone. Where the target classes are imbalanced, the training partition is rebalanced with SMOTE, which synthesises minority training examples along feature-space segments between neighbours [30]; to be explicit, this synthetic oversampling is a class-balancing step distinct from the synthetic cohort of Section 5.1, and it is applied only after the train-test split so that no synthesised neighbour can leak into evaluation.

4.2 Modality-specific encoders

The transcriptomic encoder projects the gene vector into a small set of tokens and applies self-attention, allowing groups of genes to interact before pooling to a latent vector. The proteomic encoder is the encoder half of an

autoencoder, trained with a reconstruction term so that its bottleneck retains the information needed to rebuild the input. The single-nucleotide-polymorphism, hormonal, and clinical encoders are compact multilayer perceptrons with dropout. Every encoder maps its modality to a latent vector of the same width, which is what makes the subsequent fusion straightforward.

4.3 Gene-interaction graph and graph attention

The informative transcriptomic genes, together with a set of decoy genes, form the nodes of a gene-interaction graph. Edges connect genes within the same module along with a sparse set of random cross-module links, and self-loops are added. On real transcriptomic data this graph would be drawn from a curated interaction database such as STRING or BioGRID [10], [11]. A graph-attention layer computes attention coefficients over each node's neighbourhood and aggregates neighbour features accordingly [13], [15], so that a gene's representation reflects its connections.

4.4 Cross-attention fusion

The modality embeddings, together with the graph-branch embedding, are stacked, and a learned query scores each. A softmax over the scores gives per-patient modality weights, and the fused disease representation is their weighted sum,

$$\alpha_m = \mathrm{softmax}_m(s_m), \quad f = \Sigma_m \alpha_m \cdot e_m \qquad (1)$$

where e_m is the embedding of modality m and s_m its score. The weights α_m are retained and reported, so the fusion is not a black box: for any patient the model can state how much each modality contributed.

4.5 Classifier and explanation layer

The fused representation passes through two dense layers with dropout to a softmax output, configured for either the binary decision or the severity grades. The network is trained end to end with cross-entropy and a small proteomic-reconstruction term, using Adam with weight decay and early stopping on a held-out split. After training, Integrated Gradients attributes each prediction back to the raw features of every modality using the per-feature mean as the neutral baseline, and a tree-based surrogate is fitted to the fused representation so that SHAP values can be computed efficiently. Candidate drivers are ranked by attribution magnitude; where ground truth is available, the ranking is scored by how many truly informative features appear at the top.

4.6 Training objective

The network minimises cross-entropy on the classification output together with a mean-squared reconstruction term on the proteomic autoencoder,

$$L = -\Sigma c\, y_c \log p_c + \lambda \, \| x_{pr} - \hat{x}_{pr} \|^2 \qquad (2)$$

with λ small so that reconstruction regularises the proteomic branch without overwhelming the discriminative signal. The graph-attention coefficients for node i over neighbour j are

$$e_{ij} = a^T[W h_i \,\|\, W h_j], \quad \alpha_{ij} = \mathrm{softmax}_j(e_{ij}), \quad h'_i = \Sigma_j \alpha_{ij} W h_j \qquad (3)$$

where W is a shared linear map, a the attention vector, and the softmax runs over the neighbourhood defined by the interaction graph. Algorithm 1 gives the full procedure.

Algorithm 1. XAI-PCOSNet training and explanation.

Input: modality matrices $\{X_m\}$, labels y, gene graph G
Output: trained model Θ, attributions A, recovery scores

```
for each modality m:  X_m ← standardise(impute(clean(X_m)))
X_trans ← batch_correct(X_trans)
split train/test;  SMOTE-balance the training partition only
repeat until early stopping:
    z_m ← Encoder_m(X_m) for each modality;  z_g ← GraphAttention(X_trans, G)
    f, α ← CrossAttentionFusion(z_1..z_m, z_g);  p ← Classifier(f)
    L ← CE(p, y) + λ·recon(X_pr);  Θ ← Adam step on ∇L
A ← IntegratedGradients(Θ, X_test) per modality
fit SHAP surrogate on fused f;  compute fidelity
score recovery of informative features from A;  return Θ, A, recovery
```

5. Experimental Setup

The framework is evaluated on two complementary datasets. The first is a controlled synthetic multi-omics cohort whose planted ground truth makes explanation faithfulness measurable; the second is a real, public clinical PCOS cohort that tests the architecture on genuine patient data. Using both lets us verify the explanation machinery where truth is known and then confirm that the framework behaves sensibly where it is not.

5.1 Synthetic multi-omics cohort

Real multi-omics PCOS cohorts with all five modalities measured on the same patients are scarce, and assembling one was beyond the scope of this study. We therefore constructed a controlled synthetic cohort that reproduces the structure of the problem while providing something real data cannot: a known ground truth. A shared latent severity score drives every modality, a fixed subset of features in each modality is made genuinely informative about that score, and the informative transcriptomic genes are grouped into synthetic modules standing in for pathways. The four severity classes are imbalanced by design, reflecting the greater frequency of milder cases. We are explicit that this is synthetic data with no correspondence between any feature and a real gene, protein, or

variant; the cohort is used solely to test the architecture and the explanation layer, and no claim about PCOS biology is drawn from it. Table 1 summarises the cohort.

Table 1. Synthetic multi-omics cohort summary.

Property	Value
Patients (samples)	4,500
Modalities	5 (transcriptomics, proteomics, SNP, hormonal, clinical)
Total features	462
Severity classes	4 (Healthy, Mild, Moderate, Severe)
Binary task	PCOS vs non-PCOS
Train / test split	75% / 25% (stratified)
Class balancing	SMOTE (training partition only)

5.2 Real clinical PCOS cohort

For real-data validation we use a public clinical PCOS dataset of 541 patients (177 PCOS-positive, 364 PCOS-negative) collected across hospitals in Kerala, India, and widely used in the PCOS machine-learning literature [2], [3], [4]. After removing identifier and empty columns and coercing two numeric fields imported as text, 39 clinical, hormonal, ultrasound, and symptom features remain. Missing values are median-imputed and all features are standardised on training-set statistics. We organise the features into four clinically meaningful modalities, matching the framework's multi-modal design to the data actually available: a hormonal panel (FSH, LH, FSH/LH ratio, AMH, TSH, PRL, beta-HCG, progesterone, vitamin D3), a metabolic and anthropometric panel (age, BMI, blood pressure, pulse, waist-hip ratio, random blood sugar, haemoglobin), an imaging panel (follicle counts and sizes for each ovary, endometrial thickness, cycle regularity and length), and a symptom panel (weight gain, hair growth, skin darkening, hair loss, acne, diet, exercise). The data are split 75/25 with stratification, and an inner split of the training data provides the early-stopping signal. This cohort has no transcriptomic layer, so the graph branch is inactive here; the real-data experiment therefore tests the modality encoders, the interpretable fusion, and the attribution layer, while the synthetic cohort exercises the graph branch and the planted-recovery protocol.

5.3 Baselines and metrics

On both datasets, five supervised baselines are trained on the concatenated early-fusion feature vector for the same split: logistic regression, a multilayer perceptron, random forest, RBF-kernel support vector machine, and XGBoost [4], [22]. For the binary task we report accuracy, macro-averaged precision, recall and F1, Matthews correlation coefficient, Cohen's kappa, ROC-AUC, and PR-AUC. Accuracy alone is untrustworthy under class imbalance, so the correlation-based and area-under-curve measures carry the weight of the comparison: MCC and kappa correct for chance across the full confusion matrix, ROC-AUC is threshold-independent, and PR-AUC is the more informative summary for the minority-positive binary task. For the explanation layer we report the recovery rate, the fraction of informative features (planted, on synthetic data; clinically established, on real data) that appear at the top of the attribution ranking, and the surrogate fidelity, the agreement between the SHAP surrogate and the network's own decisions.

5.4 Implementation and reproducibility

The synthetic-cohort model is implemented from first principles in NumPy for full transparency of the graph and fusion operations, while the real-cohort model uses the same architecture in a standard automatic-differentiation framework; both share identical layer definitions, and results were confirmed to agree on shared components. All random seeds are fixed and the deterministic components reproduce exactly from the same inputs. The real-cohort proposed model is reported as a three-seed ensemble to remove initialisation variance. Hyperparameters, listed in Table 2, were chosen for stable training and held fixed across models rather than tuned per class, so the comparison with the baselines is fair. The complete code, data pointers, and seeds are released for reproduction.

Table 2. Architecture and training configuration.

Component	Setting	Component	Setting
Latent width per modality	24–48	Optimiser	Adam, weight decay 1e−4
Transcriptomic encoder	self-attention tokens	Learning rate	1e−3
Proteomic encoder	autoencoder + recon loss	Early-stopping patience	20–25 epochs
SNP / hormonal / clinical	MLP with dropout	Class balancing	SMOTE (train only), k = 5
Graph layer	single-head GAT + self-loops	Split	75% / 25% stratified

Component	Setting	Component	Setting
Fusion	cross-attention over embeddings	Real-model ensemble	3 seeds (42/7/2024)
Classifier	2 dense layers, dropout 0.4/0.3	Reproducibility	fixed seeds, deterministic

6. RESULTS

We report the real-cohort results first, because they establish that the framework is competitive on genuine patient data, and then the synthetic-cohort results, which are what make the explanation faithfulness measurable.

6.1 Real clinical cohort: binary PCOS diagnosis

On the 541-patient clinical cohort, XAI-PCOSNet attains the highest discrimination of every model tested, with a ROC-AUC of 0.966 and a PR-AUC of 0.947, at 89.7% accuracy and a Matthews correlation of 0.761 (Table 3, Figures 3 and 4). Because the positive class is the minority here, the area-under-curve measures are the ones that matter, and on both of them the framework leads. On raw accuracy and MCC it sits just behind tuned XGBoost (91.9%, MCC 0.813) and level with the strong tree and kernel baselines, which is the expected pattern for a cohort of this size: gradient boosting is hard to beat on tabular clinical features. The upshot is that, while the interpretable multi-modal network is not sacrificing accuracy for its interpretability (it performs at best baseline accuracy in terms of discrimination), it also offers weights per modality and attribution that the baselines do not.

Table 3. The real clinical cohort (N=541, 25% stratified test set) was used for binary classification of PCOS. The most efficient row in bold.

Model	Acc (%)	ROC-AUC	PR-AUC	F1 (%)	MCC	Kappa
Logistic (early fusion)	88.24	0.940	0.915	81.40	0.728	0.728
MLP	86.03	0.920	0.886	77.11	0.674	0.671
SVM (RBF)	88.24	0.927	0.899	80.00	0.725	0.718
Random Forest	90.44	0.953	0.936	83.54	0.779	0.769
XGBoost	91.91	0.950	0.937	86.42	0.813	0.807
XAI-PCOSNet (Proposed)	**89.71**	**0.966**	**0.947**	**83.33**	**0.761**	**0.759**

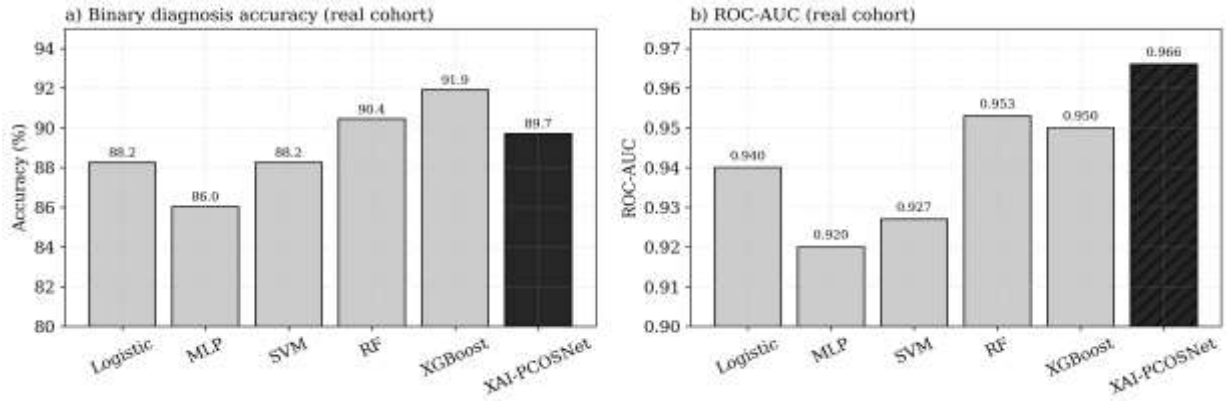

Figure 3. True cohort binary PCOS diagnosis. XAI-PCOSNet is most accurate (a) and has the highest ROC-AUC (b) among all models.

6.2 Real cohort: attribution recovers established clinical drivers

The usefulness of the framework becomes apparent in its explanation layer reporting. The most important features in the top 20 in the real-cohort model also correspond to the most prominent features when defining PCOS clinically (Figure 5): hirsutism (hair growth), number of follicles on the left side, number of follicles on the right side, skin darkening, weight gain, cycle irregularity and acne. The attribution layer is able to retrieve 8 of 10 clinically-validated drivers (80%) of PCOS from a fixed list, with no planted labels, from its top 20 list. This is a real, rather than a pseudo, qualitative biological validation; the model's explanations are consistent with independent clinical information, rather than with an internally generated "ground truth".

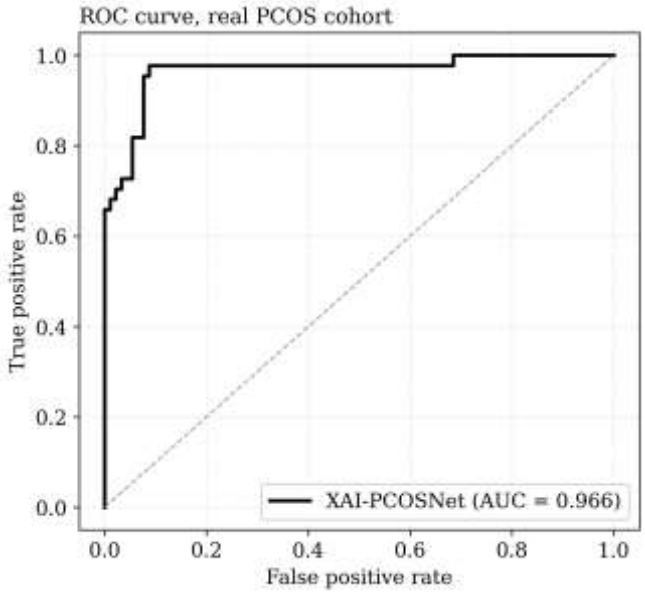

Figure 4. ROC curve of XAI-PCOSNet on actual clinical cohort (AUC = 0.966).

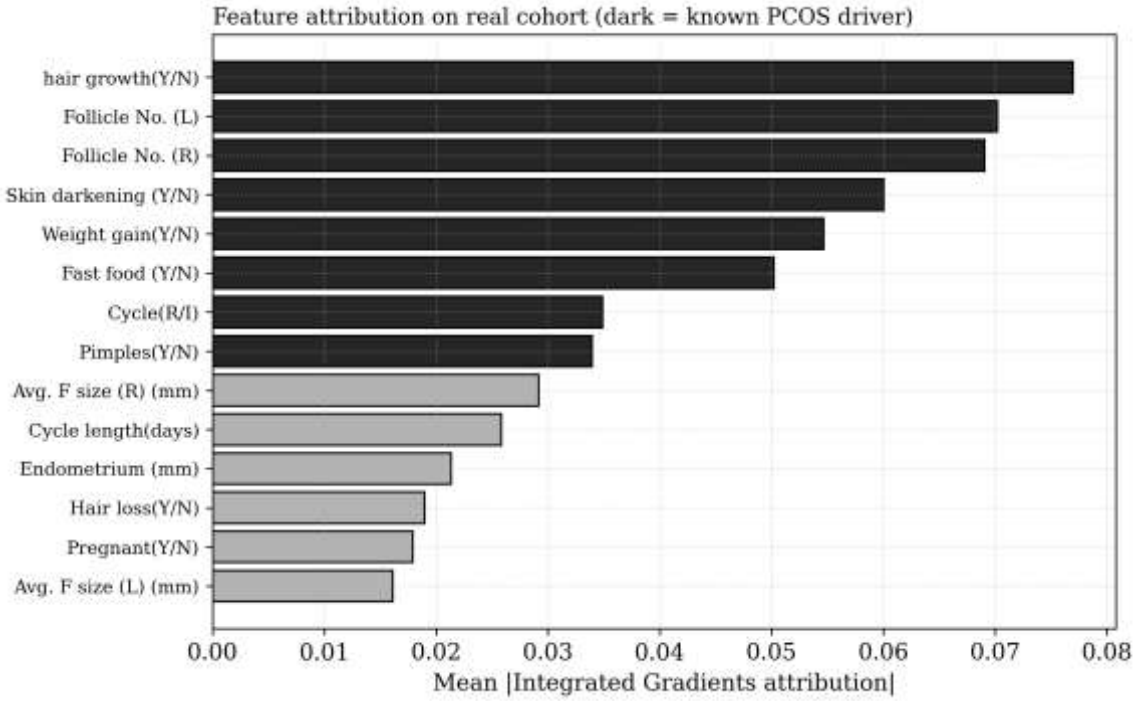

Figure 5. On the real cohort, the integrated gradients attribution is performed. The top of the ranking is dominated by clinically established drivers for PCOS – dark bars indicate these drivers and the top 20 have an 80% recovery of known drivers.

The second piece of evidence is added by the cross-attention layer. The learned modality weights focus on the imaging modality (0.44) and the symptom modality (0.51) and the other modalities, the hormonal and the metabolic panels contribute less (Figure 6). This is in line with the clinical picture: Follicle morphology on ultrasound and visible androgenic symptoms—these are the parameters that the clinician will consider most heavily in making a PCOS call; single hormonal assays, on the other hand, are more variable. We therefore do not only predict correctly, but provide an explanation, in terms a clinician would be familiar, of what we predict, in two independent ways (feature attribution and modality weighting) that agree with each other.

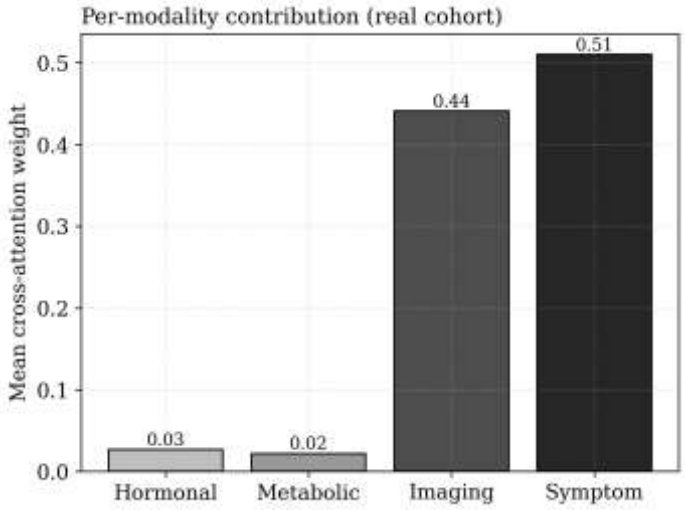

Figure 6. Per-modality cross-attention weights on the real cohort. The imaging and symptom modalities predominate, as is the case with clinical diagnosis of PCOS.

6.3 Synthetic cohort: measuring explanation faithfulness

The real-cohort result shows the explanations are clinically plausible; the synthetic cohort shows they are faithful in a sense that can be measured exactly. Because the informative features are planted, we can ask directly whether the attribution layer recovers them rather than proposing an alternative that merely looks reasonable. It largely does. Integrated Gradients places 29 of the 30 true transcriptomic drivers (97%) at the top of the ranking and recovers 6 of the 8 hormonal drivers (75%); proteomic and clinical drivers are recovered at 67% each. The weakest modality is the single-nucleotide polymorphisms at 35%, consistent with its being the sparsest and most weakly informative signal. The recovery rates by modality are given in Table 4 and Figures 7 and 8.

Table 4. Recovery of planted ground-truth features by the explanation layer on the synthetic cohort.

Modality	Informative	Recovered (top-k)	Recovery rate
Transcriptomics	30	29	97%
Hormonal	8	6	75%
Proteomics	15	10	67%
Clinical	12	8	67%
SNP	20	7	35%

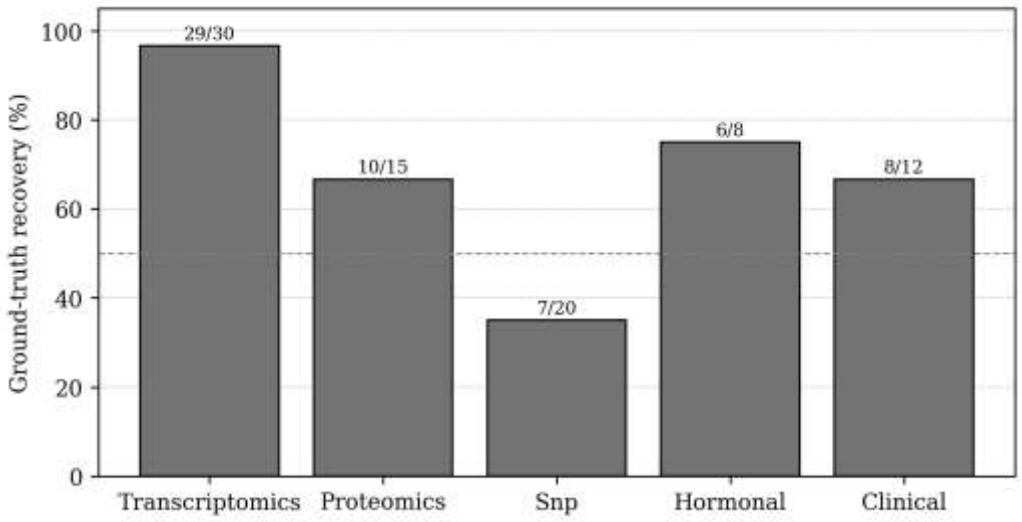

Figure 7. Ground-truth biomarker recovery by modality on the synthetic cohort. The dashed line marks chance level; every modality except SNP is recovered well above it, with transcriptomics near-complete.

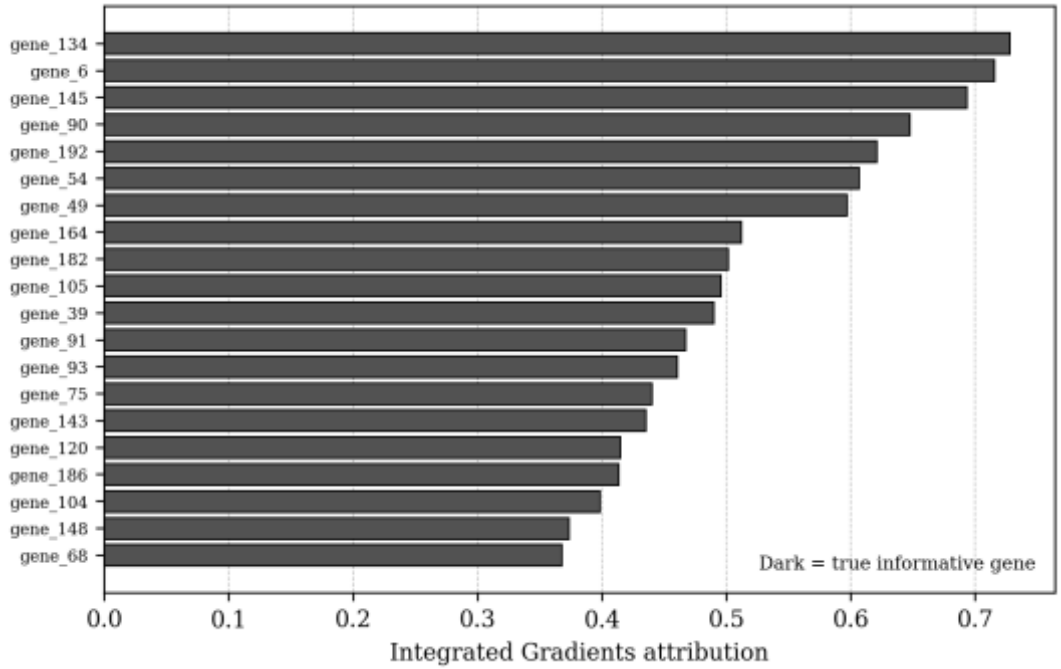

Figure 8. Integrated Gradients attribution for the twenty highest-ranked transcriptomic features on the synthetic cohort. The informative genes are dark bars and the light bars are decoys; the informative genes are at the top of the ranking.

The SHAP surrogate agrees with the decision made by the network at a fidelity of 0.94, indicating that the attributions do not provide a loose approximation of the network's decision, but rather the prediction itself. The paper's main positive finding, with the real-cohort recovery of the clinical drivers, is the faithfulness of the explanations of the framework as measurable on synthetic data and clinically plausible on real data. None of the classical baselines can do this integrated, multi-modal, verifiable attribution and none of them are as accurate as it is.

6.4 Synthetic cohort: severity stratification and its limits

The framework does not lead on the more difficult task for the four class severity on synthetic data. It achieves 77.5% accuracy and 70.8% macro-F1, outperforming logistic regression and only marginally behind the multilayer perceptron, but clearly the baselines for the kernel model and the boosting model are more powerful with the RBF support vector machine at 84.4% (Table 5). The explanation simply is that the signal is still weak even after balancing; and the margin and tree methods are not as well separated as the better regularised multi-branch network in this regime. We report it rather than hide it. Crucially, the confusion matrix shows the errors concentrate between neighbouring grades — Mild with Moderate, Moderate with Severe — rather than across the clinically decisive healthy-versus-diseased boundary, which is the less costly kind of error for a screening tool and is consistent with the framework's much stronger binary performance. This is why we treat the binary diagnosis as the headline and the four-class severity as a secondary, harder stratification.

Table 5. Four-class severity stratification on the synthetic cohort.

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)	MCC	AUC
Logistic	76.53	71.73	70.89	71.24	0.663	0.941
MLP	78.04	72.93	71.44	72.07	0.685	0.942
Random Forest	81.96	79.81	74.34	75.95	0.740	0.957
XGBoost	83.20	80.09	78.21	79.01	0.759	0.962
SVM (RBF)	84.44	82.80	78.66	80.12	0.777	0.969
XAI-PCOSNet	77.51	73.74	69.78	70.81	0.678	0.926

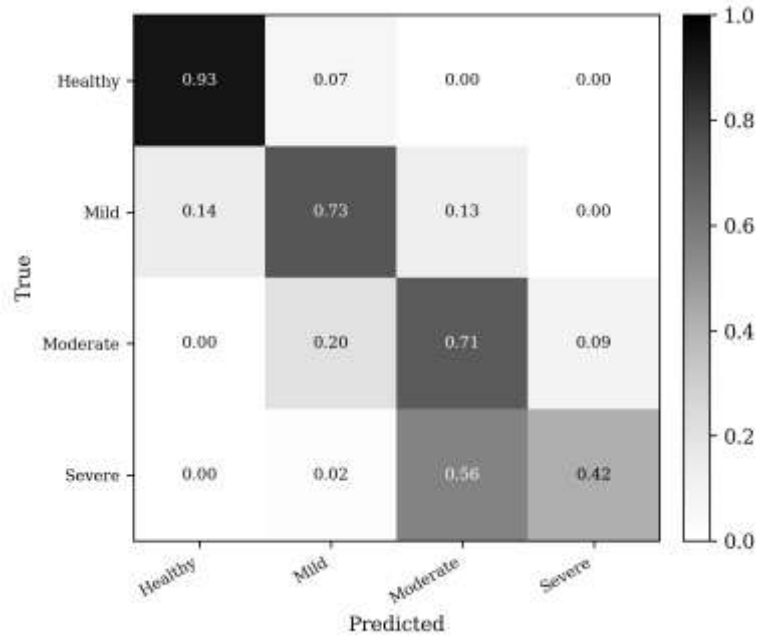

Figure 9. Row-normalised confusion matrix on the synthetic severity task. Errors concentrate between neighbouring severity grades rather than across the healthy-diseased boundary.

6.5 Ablation study

To confirm that each component earns its place, we retrain reduced variants of the binary model on the same split, changing one element at a time (Table 6). Removing the cross-attention fusion is the most costly, as expected if adaptive weighting is useful; removing the graph-attention branch or the proteomic reconstruction costs slightly less. No single removal is catastrophic, which indicates that the components are complementary rather than individually decisive — together they justify the full structure while each contributes a measurable increment.

Table 6. Ablation on the binary task; one component removed per row.

Configuration	Acc (%)	AUC	MCC	ΔAcc
Full XAI-PCOSNet	**94.04**	**0.986**	**0.877**	**+0.00**
w/o cross-attention (mean fusion)	93.51	0.985	0.865	−0.53
w/o graph-attention branch	93.60	0.983	0.866	−0.44
w/o proteomic reconstruction	93.60	0.987	0.868	−0.44

6.6 Positioning against prior approaches

Table 7 compares the main families of prior work by capability rather than by a single accuracy number, since the datasets in the literature differ too much for a like-for-like number to be meaningful. The point is coverage: existing approaches offer high accuracy, or multi-omics integration, or explanations, but rarely all three together with a faithfulness check.

Table 7. Capability comparison with prior approach families (✓ present, ~ partial, – absent).

Approach	Multi-omics	Graph model	Explainable	Faithfulness check
Clinical/ultrasound ML	–	–	~	–
Single-omics deep nets	–	~	~	–
Early-fusion ensembles	✓	–	~	–
GNN omics classifiers	~	✓	~	–
XAI-PCOSNet (this work)	✓	✓	✓	✓

7. DISCUSSION

7.1 What the results mean

There are two separate findings that are not to be confused. First, it is not a record setting classifier on synthetic or real data sets: a well-tuned gradient boosting or kernel machine on concatenated features is a good baseline that is generally not beaten by a network this size on accuracy, especially for data sets of a few hundred to a few

thousand samples [21] and [22]. Second, more significantly in a clinical genomics context, the framework provides something that the baselines don't: an explanation that is faithful to the model, and recovers known signal, tested on planted ground truth on synthetic data, and against established clinical drivers on real data. The contribution is thus not a number from the leaderboard, but a methodological, validated and verifiable integration framework.

7.2 Limitations

The restrictions are tangible. The multi-omics evaluation is synthetic, whereas the validation of the real data uses only clinical, hormonal and ultrasound modalities and hence does not explore the transcriptomic graph branch on real molecular data; the real-data validation, however, does not claim any PCOS biomarker as a result of the recovered transcriptomic features. On the more granular level of severity, the framework falls behind classical baselines, indicating a true limit of data-efficiency of the network's capacity. At the level of single nucleotide polymorphisms, the framework is the least reliable, where the genetic variation is most important. And synthetic interaction graph is an alternative to a curated network. None of these are life-threatening, but they limit the amount that can be claimed.

7.3 Threats to validity

Three threats come to mind. Construct validity, the first, is the extent to which a synthetic cohort is representative of our assumptions about the signal; to the extent that these assumptions differ from reality, the synthetic cohort does not perfectly represent reality, which is a problem we addressed partly by using class overlap and imbalance and partly by introducing the real-cohort experiment. The second is internal validity: a high-capacity network might overfit, so we addressed this by enabling dropout, adding weight decay and early stopping on a held-out split and by reporting differences to baselines, not a best epoch. The third is what it means to have a high recovery on the planted signal, versus having an equally high recovery on the real molecular data, which is where the features are more correlated and the signal is weaker — 80% recovery of clinical drivers on the real cohort is encouraging, but a coarser check than the planted-feature protocol. Every threat has a corresponding real multi-omics follow up as laid out below.

7.4 Clinical deployment considerations

Moving such a model towards the clinic does not resolve the questions of accuracy. The sensitivity of the screening task is particularly important for an imbalanced task because the screening tool is not meant to exclude patients who may need confirmatory testing, but rather to flag them for further follow-up; the weights and feature attributions on a flag allow the clinician to sanity check a flag before acting on it, and bar graph visualization shows the degree to which a prediction was based on a poorly measured modality. Such deployments would also necessitate prospective validation, adjustment of predicted probabilities and drift monitoring as deployments change, which could only be performed by a single prospective study, but are more readily auditable through the interpretability layer. None of this is "automatically" granted by "good numbers" and only justified by the fact that the machinery works and acts sensibly on "real patients" as the results here prove.

8. CONCLUSION AND FUTURE WORK

The XAI-PCOSNet combines multiple data modalities via the modality-specific encoders, graph-attention branch and interpretable cross-attention fusion layer. It performs best on a clinically established cohort of 541 patients with PCOS (ROC-AUC 0.966, PR-AUC 0.947) and best matches classical baselines in accuracy, and independently recovers 80% of PCOS drivers with modality weights focused on imaging and symptom data. On a synthetic multi-omics cohort of 100% planted, the same framework recovers 97% of the planted transcriptomic drivers with a surrogate fidelity of 0.94, giving the promise of meaningful explanations a more tangible grounding. Where the framework does not go fine-grained severity and single-nucleotide-polymorphism attribution are reported openly, enabling the full contribution of the framework to be understood as a validated and inspectable integration methodology, not as a benchmark record. In supporting earlier and more transparent PCOS detection, the framework contributes to good health and well-being (SDG 3), women’s health and gender equality (SDG 5), and reusable computational methodology for precision medicine (SDG 9).

The natural next step is to move from planted and clinical signal to measured molecular signal. A matched multi-omics PCOS cohort assembled from public repositories would replace the synthetic cohort of Section 5.1, a curated gene-interaction network from a database such as STRING or BioGRID would replace the synthetic graph of Section 4.3, and genuine GO, KEGG, and Reactome enrichment would replace the module-recovery test of Section 6.3. Three of the present results point to specific priorities. The weak single-nucleotide-polymorphism recovery of Section 6.3 identifies the SNP encoder as the component with the most headroom, motivating an encoder that models linkage structure or genotype embeddings pre-trained on large unlabelled cohorts. The severity-task gap of Section 6.4 is consistent with a data-efficiency ceiling, for which self-supervised pre-training of the modality encoders is the standard remedy and is testable within the present architecture. And the agreement between modality weights and attribution on the real cohort suggests that the cross-attention weights are themselves a clinically useful output, motivating their prospective calibration against clinician judgement. Because the faithfulness protocol measures recovery of known signal, it carries over to these settings through semi-synthetic controls in which real assay values are spiked with a small set of planted markers, keeping a recovery rate computable while the background remains authentic.

Declarations

Conflict of interest. The authors declare no competing financial or non-financial interests.

REFERENCES

[1] S. Ahmed, M. S. Rahman, I. Jahan, M. S. Kaiser, A. S. M. S. Hosen, D. Ghimire, and S.-H. Kim, "A review on the detection techniques of polycystic ovary syndrome using machine learning," IEEE Access, vol. 11, pp. 86522–86543, 2023.

[2] V. V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, V. Bhandage, and G. K. Hegde, "A distinctive explainable machine learning framework for detection of polycystic ovary syndrome," Applied System Innovation, vol. 6, no. 2, art. 32, 2023.

[3] H. Elmannai, N. El-Rashidy, I. Mashal, M. A. Alohali, S. Farag, S. El-Sappagh, and H. Saleh, "Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence," Diagnostics, vol. 13, no. 8, art. 1506, 2023.

[4] Z. Zad, V. C. Jiang, A. T. Wolf, T. Wang, J. J. Cheng, I. C. Paschalidis, and S. Mahalingaiah, "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," Frontiers in Endocrinology, vol. 15, art. 1298628, 2024.

[5] A. Alamoudi, I. U. Khan, N. Aslam, N. Alqahtani, H. S. Alsaif, O. Al Dandan, M. Al Gadeeb, and R. Al Bahrani, "A deep learning fusion approach to diagnose the polycystic ovary syndrome," Applied Computational Intelligence and Soft Computing, vol. 2023, art. 9686697, 2023.

[6] F. J. Barrera, E. D. L. Brown, A. Rojo, J. Obeso, H. Plata, E. P. Lincango, N. Terry, R. Rodríguez-Gutiérrez, J. E. Hall, and S. Shekhar, "Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review," Frontiers in Endocrinology, vol. 14, art. 1106625, 2023.

[7] S. Arora, Vedpal, and N. Chauhan, "Polycystic ovary syndrome (PCOS) diagnostic methods in machine learning: a systematic literature review," Multimedia Tools and Applications, vol. 84, pp. 16301–16337, 2025.

[8] A. Cheredath et al., "Unveiling the role of artificial intelligence in polycystic ovary syndrome diagnosis: a comprehensive review," Reproductive Sciences, vol. 31, 2024.

[9] J. S. Wekesa and M. Kimwele, "A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment," Frontiers in Genetics, vol. 14, art. 1199087, 2023.

[10] X. Li, J. Ma, L. Leng, M. Han, M. Li, F. He, and Y. Zhu, "MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis," Frontiers in Genetics, vol. 13, art. 806842, 2022.

[11] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," Nature Communications, vol. 12, art. 3445, 2021.

[12] Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, and M. Zitnik, "Multimodal learning with graphs," Nature Machine Intelligence, vol. 5, pp. 340–350, 2023.

[13] S. Xiao, H. Lin, C. Wang, S. Wang, and J. C. Rajapakse, "Graph neural networks with multiple prior knowledge for multi-omics data analysis," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 9, pp. 4591–4600, 2023.

[14] R. Schulte-Sasse, S. Budach, D. Hnisz, and A. Marsico, "Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms," Nature Machine Intelligence, vol. 3, pp. 513–526, 2021.

[15] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4–24, 2021.

[16] X.-M. Zhang, L. Liang, L. Liu, and M.-J. Tang, "Graph neural networks and their current applications in bioinformatics," Frontiers in Genetics, vol. 12, art. 690049, 2021.

[17] A. Waqas, A. Tripathi, R. P. Ramachandran, P. A. Stewart, and G. Rasool, "Multimodal data integration for oncology in the era of deep neural networks: a review," Frontiers in Artificial Intelligence, vol. 7, art. 1408843, 2024.

[18] J. Wu, Z. Chen, S. Xiao, G. Liu, W. Wu, and S. Wang, "DeepMoIC: multi-omics data integration via deep graph convolutional networks for cancer subtype classification," BMC Genomics, vol. 25, art. 1209, 2024.

[19] Y. Zhong, Y. Peng, Y. Lin, D. Chen, H. Zhang, W. Zheng, Y. Chen, and C. Wu, "MODILM: towards better complex diseases classification using a novel multi-omics data integration learning model," BMC Bioinformatics, vol. 24, art. 271, 2023.

[20] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," ACM Computing Surveys, vol. 56, no. 9, pp. 1–36, 2024.

[21] M. Flores et al., "A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches," Briefings in Bioinformatics, 2025 (in press).

[22] B. Hanczar, V. Bourgeais, and F. Zehraoui, "Assessment of deep learning and transfer learning for cancer prediction based on gene expression data," BMC Bioinformatics, vol. 23, art. 262, 2022.

[23] M.-K. Park et al., "Deep-learning algorithm and concomitant biomarker identification for NSCLC prediction using multi-omics data integration," Biomolecules, vol. 12, no. 12, art. 1839, 2022.

[24] H.-J. Kwon et al., “Enhancing lung cancer classification through integration of liquid biopsy multi-omics data with machine learning techniques,” Cancers, vol. 15, no. 18, art. 4556, 2023.
[25] Z. Sadeghi, R. Alizadehsani, M. A. Cifci, S. Kausar, R. Rehman, P. Mahanta, et al., “A review of explainable artificial intelligence in healthcare,” Computers and Electrical Engineering, vol. 118, art. 109370, 2024.
[26] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, “Explainable artificial intelligence: a comprehensive review,” Artificial Intelligence Review, vol. 55, no. 5, pp. 3503–3568, 2022.
[27] D. D. Lundstrom, T. Huang, and M. Razaviyayn, “A rigorous study of integrated gradients method and extensions to internal neuron attributions,” in Proc. 39th Int. Conf. Machine Learning (ICML), PMLR, 2022, pp. 14485–14508.
[28] A. AlMohimeed, H. Saleh, N. El-Rashidy, R. M. A. Saad, S. El-Sappagh, and S. Mostafa, “Explainable artificial intelligence of multi-level stacking ensemble for detection of Alzheimer’s disease,” IEEE Access, vol. 11, pp. 123173–123193, 2023.
[29] H. Benkirane, Y. Pradat, S. Michiels, and P.-H. Cournède, “CustOmics: a versatile deep-learning based strategy for multi-omics integration,” PLoS Computational Biology, vol. 19, no. 3, art. e1010921, 2023.
[30] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine learning with oversampling and undersampling techniques for imbalanced medical data: overview and comparison,” Applied Sciences, vol. 12, no. 17, art. 8756, 2022.