Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 22 July 2026
Revised 28 Aug 2026
Accepted 10 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (22s): 64–91

KEYWORDS:
Polycystic Ovary Syndrome; Temporal Deep Learning; Federated Learning; Uncertainty Estimation; Ultrasound Classification; Offline Reinforcement Learning

A Modular Causal-Structure-Guided Temporal and Federated Deep Learning Framework for PCOS Prediction and Offline Intervention Ranking

Priya Raghunath Jamdade[1], Waseem Ahmad Mir[2], Poonam Balasaheb Dhole[3]

Department of Computer Science and Engineering, School of Computational Sciences, JSPM University, Pune, India
jamdadepriya1993@gmail.com, waseemmir78177@gmail.com, Poonam.dhole2218@gmail.com

ABSTRACT: Polycystic Ovary Syndrome (PCOS) is a heterogeneous endocrine and metabolic disorder for which computational assessment may benefit from complementary clinical and imaging analyses. This study presents a modular clinical-dependency-guided temporal and federated deep learning framework comprising a Temporal-Aware Self-Supervised Contrastive Transformer (TASS-Transformer), a Clinical Graph-Augmented Variational Bayesian Network (CG-VBN), a Federated Ensemble Attention-CNN (FEA-CNN), a Clinically Informed Feature Denoising Autoencoder (CIF-DAE), and an offline Adaptive Intervention Recommender (AIR-RL). The clinical branch was evaluated on 541 unique public records (177 PCOS-positive and 364 non-PCOS), while the independent ultrasound branch used 1,924 images from a separate public repository. The repositories are not patient-linked; therefore, the branches were evaluated independently and no end-to-end patient-level multimodal metric is reported. Synthetic 24-step sequences were generated only inside clinical development folds, five non-IID ultrasound clients were simulated, and AIR-RL used simulated transitions rather than observed treatment trajectories. Repeated internal validation produced 98.31% accuracy, 97.63% F1-score, 99.12% AUC, and 98.14% specificity for the clinical branch; the ultrasound module achieved 98.11% image-level accuracy. AIR-RL achieved 95.61% simulated intervention success and 96.82% safety-rule compliance. These findings demonstrate internal computational feasibility only. No external cohort, real hospital federation, formal differential-privacy mechanism, causal treatment effect, or prospective treatment outcome was evaluated. Patient-linked longitudinal multicenter validation is required before clinical use

INTRODUCTION

PCOS is one of the most prevalent endocrine and metabolic disorders affecting women of reproductive age [1–3]. It is characterized by considerable heterogeneity in reproductive, hormonal, metabolic, and psychological manifestations, including menstrual irregularity, hyperandrogenism, insulin resistance, obesity, infertility, and an increased risk of long-term cardiometabolic complications. Despite its substantial clinical burden, PCOS diagnosis continues to rely mainly on rule-based frameworks, including the Rotterdam criteria and related clinical guidelines [4–6]. These criteria are valuable for standardized diagnosis but may not fully represent the complex interactions among endocrine markers, metabolic abnormalities, lifestyle factors, and ovarian morphology.

Machine learning has increasingly been applied to PCOS prediction using clinical, biochemical, demographic, and

ultrasonographic attributes. Conventional classifiers can identify nonlinear associations and improve automated screening; however, most existing approaches treat clinical variables as static and independent observations. Consequently, they provide limited representation of cyclic endocrine variation, uncertainty in incomplete records, and dependencies among reproductive and metabolic factors. Deep learning methods offer greater capacity for extracting complex representations from clinical and imaging data [7–9], but their application is frequently limited by inadequate interpretability, sensitivity to dataset shift, small single-source datasets, and difficulty in transferring models across heterogeneous healthcare environments. Recent studies have also examined electronic health record-based prediction, explainable machine learning, feature selection, and optimized diagnostic models [10–18]. Nevertheless, temporal modeling, calibrated uncertainty estimation, privacy-aware distributed learning, and intervention-oriented decision support are generally investigated as separate problems.

Advances in temporal transformers, graph-based probabilistic learning, federated learning, clinically informed dependency-structure representation learning, and reinforcement learning provide opportunities to address these limitations. Temporal transformers can model repeated or cyclic patterns in endocrine measurements, while graph-based Bayesian methods can represent dependencies among clinical variables and quantify predictive uncertainty. Federated learning offers a mechanism for investigating distributed model training without centralizing source data, although its effectiveness must be evaluated carefully under controlled heterogeneous-client settings. Similarly, reinforcement learning may support the offline ranking of candidate management options, but simulated policy performance cannot be interpreted as evidence of clinical treatment effectiveness without prospective intervention data.

PCOS is not adequately represented as a fixed and uniform condition because its manifestations vary across individuals and may change over time. Static models may therefore overlook interactions among hormone-related variables, metabolic indicators, menstrual characteristics, lifestyle factors, and ovarian imaging findings. Moreover, systems designed only for binary classification provide limited support for uncertainty-aware interpretation or subsequent management planning. These limitations motivate a modular framework that examines temporal representation, clinical dependency modeling, image classification, feature denoising, uncertainty estimation, and safety-constrained intervention ranking within clearly defined experimental boundaries.

Accordingly, this study proposes a modular clinical-dependency-guided temporal and federated deep learning framework comprising five components: a Temporal-Aware Self-Supervised Contrastive Transformer for training-only synthetic temporal representation; a Clinical Graph-Augmented Variational Bayesian Network for uncertainty-aware clinical inference; a Federated Ensemble Attention-CNN for ovarian ultrasound classification under simulated non-identically distributed clients; a Clinically Informed Feature Denoising Autoencoder for refining clinical and lifestyle representations; and an offline Adaptive Intervention Recommender for ranking candidate management options under predefined safety constraints. Here, clinically informed dependency guidance denotes the use of domain-informed structural constraints for representation learning; no causal treatment effect, counterfactual effect, or treatment-response parameter is estimated from the observational dataset.

The clinical and ultrasound repositories are independent and are not patient-linked. Accordingly, the clinical branch and image branch are trained and evaluated separately, and no end-to-end patient-level multimodal performance metric is reported. The fusion operator is retained only as a future architecture for prospectively linked records in which clinical and ultrasound observations are verified to belong to the same patient within a defined time window. Synthetic temporal sequences are generated only within development folds, the federated clients are simulated rather than real hospitals, and AIR-RL is evaluated using offline simulated transitions; no prospective treatment outcome is represented.

The principal contribution of this study is therefore a transparent modular proof of concept that integrates temporal representation learning, uncertainty-aware graph inference, simulated federated ultrasound classification, clinical-dependency-guided denoising, and offline safety-constrained intervention ranking. The study evaluates diagnostic discrimination, calibration, robustness, computational cost, federated behavior, explainability, and module-wise contributions while explicitly distinguishing original observations from augmented or simulated research inputs. The proposed framework is intended to support future development of PCOS decision-support systems and requires patient-linked multimodal datasets, independent multicenter validation, genuine longitudinal measurements, and prospective clinician-led evaluation before any clinical deployment or treatment-effect claim can be considered.

1. RELATED WORK AND CRITICAL RESEARCH GAPS

Artificial intelligence has increasingly been applied to PCOS screening, diagnosis, risk assessment, and clinical feature interpretation. Early studies primarily used structured clinical, demographic, hormonal, and metabolic variables with conventional machine learning classifiers. These investigations demonstrated that algorithms such as logistic regression, support vector machines, decision trees, random forests, and ensemble classifiers can distinguish PCOS-positive cases from controls using routinely available patient attributes [1–3]. Tong et al. [2] developed a machine learning-based diagnostic model, while Tiwari et al. [3] proposed the SPOSDS framework for automated PCOS screening. Although these studies established the feasibility of computational diagnosis, their predictions were mainly derived from static tabular observations and did not explicitly model temporal hormone variation, uncertainty, or changing clinical states.

Feature selection and classifier optimization have subsequently been investigated to improve diagnostic performance and reduce redundant clinical inputs. Danaei Mehr and Polat [17] compared different machine learning and feature-selection techniques, whereas Panjwani et al. [18] combined optimized feature selection with classifier tuning for early PCOS detection. Reka et al. [24] similarly examined a hybrid machine learning strategy based on stable predictive features. These approaches demonstrate that selecting clinically informative attributes can improve model efficiency and reduce overfitting. However, most feature-selection models treat variables as independent predictors and provide limited representation of the physiological dependencies among hormonal imbalance, insulin resistance, menstrual irregularity, obesity, and ovarian morphology.

Electronic health records and alternative physiological signals have also been explored as sources for PCOS prediction. Zad et al. [11] evaluated machine learning models using electronic health record data, demonstrating the potential of routinely collected clinical information for scalable screening. Lim et al. [20] investigated radial pulse-wave characteristics as a non-invasive physiological input for machine learning classification. Such approaches broaden the range of information available for PCOS assessment and may reduce dependence on specialized imaging. Nevertheless, electronic health records frequently contain missing, inconsistent, and institution-specific measurements, while alternative physiological signals require broader population-level validation before they can be incorporated into routine diagnostic practice.

Ultrasound imaging represents another important area of AI-based PCOS research because ovarian morphology, follicle distribution, and ovarian volume are clinically relevant diagnostic indicators. Suha and Islam [4] developed an extended machine learning approach for PCOS detection using ovarian ultrasound images. Imaging-based models can identify complex visual patterns that may be difficult to represent using manually defined features. However, many available ultrasound studies are based on relatively small, single-repository datasets and image-level labels. Furthermore, the absence of verified pixel-level masks in commonly used public datasets limits the validity of lesion-localization or segmentation claims. Attention maps such as Grad-CAM can indicate image regions that influence classification, but they should not be interpreted as quantitative lesion-localization results without corresponding ground-truth annotations.

Explainability has received increasing attention because clinical adoption requires more than high predictive accuracy. Khanna et al. [12] proposed an explainable machine learning framework for PCOS detection, while broader reviews have emphasized the need for transparent feature attribution and clinically understandable predictions [15,19]. Techniques such as SHAP can rank the relative influence of hormonal, metabolic, menstrual, and demographic variables. However, many studies present feature rankings without evaluating calibration, predictive uncertainty, explanation stability, or agreement with independent clinical experts. Explainability should therefore be considered separately from diagnostic accuracy, and percentage-based explainability scores should not be reported unless a clearly defined annotation and evaluation protocol is available.

Bioinformatics and biomarker-oriented studies have extended PCOS research beyond conventional clinical variables. Nasim et al. [16] examined machine learning in a bioinformatics context, while Li et al. [23] used machine learning to identify regulated cell-death-related genes associated with PCOS. Wang et al. [26] investigated OVGP1 as a potential diagnostic biomarker, and Xu et al. [29] analyzed transcriptomic associations between PCOS and metabolic syndrome. These studies provide important insights into molecular mechanisms and biomarker discovery. Nevertheless, genomic and transcriptomic models often require high-dimensional datasets, independent biological validation, and specialized laboratory measurements that may not be routinely available in primary healthcare settings.

Several clinical and biological reviews provide the medical foundation required to interpret AI-based PCOS models. Franks [5], Burt Solorzano and McCartney [6], Stener Victorin et al. [7], Joham et al. [9], and Barber and Franks [30] describe the endocrine,

reproductive, metabolic, and obesity-related characteristics of PCOS. Standeven et al. [8] discuss psychiatric and behavioral implications, while Louwers et al. [21], Joham et al. [22], and Dumesic et al. [28] examine genetic, epidemiological, and evolutionary perspectives. These works are important for defining clinically meaningful variables and understanding disease heterogeneity. However, they are not reproducible diagnostic algorithms and should therefore be used as contextual literature rather than numerical baselines in experimental comparisons.

Recent work has also expanded from primary diagnosis toward reproductive risk and decision-support applications. Evans et al. [25] proposed OVARIFY using clinical features for PCOS prediction, while Mogos et al. [27] applied machine learning to predict unfavorable pregnancy outcomes among patients with PCOS. These studies indicate a gradual transition from binary diagnosis toward outcome-oriented analysis. However, prediction of pregnancy risk or association with clinical outcomes is different from the optimization of treatment decisions. A model cannot claim treatment effectiveness unless recommended interventions are prospectively administered and patient outcomes are observed over time.

Federated learning may provide a useful mechanism for training models across distributed data sources without transferring raw patient records. This is particularly relevant to ultrasound imaging and clinical data that may be stored within separate institutions. However, most available PCOS datasets are public, small, and not organized as real multi-hospital federations. Consequently, federated experiments based on artificially partitioned datasets should be described as simulated non-identically distributed client experiments, rather than evidence of deployment across real hospitals. Such simulations can evaluate client drift, aggregation behavior, and privacy-preserving model exchange, but they cannot establish multicenter clinical generalizability.

Temporal modeling remains comparatively underdeveloped in PCOS research. Hormonal markers and menstrual characteristics are inherently dynamic, but most public PCOS datasets contain cross-sectional observations rather than repeated measurements collected over multiple cycles. Temporal transformers and self-supervised sequence-learning methods may capture cyclic patterns when genuine longitudinal data are available. When temporal sequences are generated from static records, however, the resulting analysis must be described as controlled temporal augmentation. Generated sequences can support methodological experimentation but cannot replace prospectively collected longitudinal hormone measurements or demonstrate early detection before clinical diagnosis.

Reinforcement learning offers another potential direction for ranking management options according to patient states, predicted risks, and predefined safety constraints. Nevertheless, reinforcement learning requires reliable trajectories containing states, actions, and subsequent outcomes. Public PCOS datasets generally do not provide longitudinal records of treatment selection and observed response. Therefore, reinforcement-learning components trained using bounded transition models should be evaluated as offline simulations. Measures such as simulated intervention success, reference-policy concordance, safety-rule compliance, and offline policy value are appropriate, whereas clinical treatment success, patient improvement, or long-term therapeutic effectiveness cannot be claimed without prospective outcome data.

Table 1 summarizes the computational, imaging, explainability, biomarker-oriented, and clinical studies relevant to PCOS prediction. The table is intended as a structured index of the reviewed literature rather than a source of directly comparable experimental results. Only studies that define reproducible algorithms and can be implemented using identical data partitions, preprocessing procedures, and evaluation metrics are eligible for quantitative benchmarking. Clinical reviews, psychiatric analyses, epidemiological investigations, genetic studies, and evolutionary theories are retained to support medical interpretation but are not treated as computational baselines.

Table 1. Comparative review of computational, imaging, explainability, and clinical studies relevant to PCOS prediction

Reference	Method	Main Objectives	Findings	Limitations
[1]	Machine Learning-Based PCOS Prediction Framework	Investigate applicability of machine learning algorithms for PCOS diagnosis.	Demonstrated feasibility of automated PCOS prediction using clinical parameters	Limited dataset size and absence of multimodal features

[2]	Machine Learning Diagnostic Model	Develop an accurate diagnostic model for PCOS detection	Achieved improved diagnostic performance through optimized feature learning	Limited explainability and temporal disease modeling
[3]	SPOSDS Smart Diagnostic System	Create an intelligent PCOS screening and decision-support system	Improved diagnostic automation and prediction accuracy	Dependent on structured clinical features only
[4]	Extended Ultrasound Image-Based Machine Learning Model	Detect PCOS using ovarian ultrasound imaging	Demonstrated effectiveness of imaging biomarkers for diagnosis	Limited integration with hormonal and metabolic factors
[5]	Clinical review of PCOS pathophysiology and manifestations	Reviewed pathophysiology and clinical manifestations of PCOS	Established endocrine and metabolic foundations of PCOS	No computational diagnostic framework
[6]	Clinical review of endocrine and metabolic dysfunction	Analyze hormonal and reproductive dysfunction in PCOS	Highlighted multifactorial disease mechanisms	Lack of predictive analytics
[7]	Comprehensive clinical disease primer	Provide integrated understanding of PCOS pathogenesis and management	Identified genetic, metabolic, and reproductive interactions	Focused on clinical understanding rather than predictive systems
[8]	Psychiatric and behavioral health review	Investigate psychological implications of PCOS	Established strong association between PCOS and mental health disorders	Absence of machine learning integration
[9]	Clinical and epidemiological review	Examine diagnosis and management strategies	Identified challenges in early diagnosis and treatment	Limited use of advanced AI methodologies
[10]	Conceptual review of AI and machine learning in PCOS	Assess AI applications in PCOS diagnosis	Demonstrated growing significance of AI-assisted healthcare solutions	Primarily conceptual analysis
[11]	EHR-Based Machine Learning Prediction Model	Predict PCOS using electronic health records	Achieved scalable diagnosis using real-world healthcare data	Susceptible to missing and noisy records
[12]	Explainable Machine Learning Framework	Improve transparency of PCOS prediction models	Enhanced interpretability using explainable AI techniques	Trade-off between explainability and complexity
[13]	Updated clinical disease review	Summarize current advances in PCOS understanding	Provided updated clinical and biological insights	No computational implementation

[14]	Clinical review of PCOS diagnosis and management	Review diagnosis and management practices	Highlighted emerging diagnostic challenges	Lack of predictive analytics framework
[15]	Review of machine learning-based PCOS detection	Surveyed machine learning approaches for PCOS detection	Identified strengths and weaknesses of existing models	Limited discussion of multimodal and federated learning
[16]	Bioinformatics-Based Machine Learning Framework	Predict PCOS using biological and clinical data	Improved prediction through bioinformatics feature integration	Limited generalization across populations
[17]	Feature Selection and Machine Learning Model	Improve diagnosis through optimal feature selection	Demonstrated importance of relevant clinical features	Limited capability for handling nonlinear temporal dependencies
[18]	Optimized Machine Learning Detection Framework	Enhance early diagnosis using optimized classifiers	Improved prediction performance for early-stage PCOS	Restricted interpretability and personalization
[19]	Systematic review of machine learning diagnostic studies	Analyze machine learning-based PCOS diagnostic methodologies	Identified trends toward hybrid and explainable models	Limited focus on intervention recommendation systems
[20]	Radial Pulse Wave Machine Learning Classifier	Diagnose PCOS using non-invasive physiological signals	Demonstrated effectiveness of pulse wave biomarkers	Limited clinical feature diversity
[21]	Review of genetic origins and susceptibility	Investigate genetic origins of PCOS	Identified genetic factors associated with disease susceptibility	No predictive clinical deployment
[22]	Epidemiological review of PCOS	Analyze prevalence and population-level impacts of PCOS	Established epidemiological burden of PCOS globally	Limited computational application
[23]	Machine Learning-Based Gene Discovery Model	Identify regulated cell death genes associated with PCOS	Discovered novel molecular biomarkers linked to PCOS	Limited clinical validation
[24]	Hybrid Machine Learning with Stable Features	Improve early diagnosis using hybrid learning methods	Enhanced robustness through stable feature extraction	Limited integration of imaging and temporal information
[25]	OVARIFY Clinical Prediction Model	Predict PCOS using machine learning and clinical features	Demonstrated clinical applicability of AI-assisted diagnosis	Limited incorporation of multimodal biomarkers
[26]	OVGP1 Biomarker Detection Framework	Diagnose PCOS using OVGP1 and machine learning algorithms	Identified OVGP1 as a promising diagnostic biomarker	Focused on a single biomarker pathway

[27]	Pregnancy Outcome Prediction Model	Predict adverse pregnancy outcomes in PCOS patients	Improved risk assessment for reproductive complications	Limited focus on primary PCOS diagnosis
[28]	Evolutionary metabolic adaptation theory	Investigate evolutionary origins of PCOS	Proposed metabolic adaptation theory explaining disease persistence	No diagnostic implementation
[29]	Transcriptomic and Machine Learning Framework	Study associations between PCOS and metabolic syndrome	Identified transcriptomic signatures linked to metabolic dysfunction	Requires high-dimensional genomic datasets
[30]	Clinical review of obesity and PCOS interactions	Examine role of obesity in PCOS progression	Established obesity as a major contributor to metabolic and reproductive abnormalities	Absence of predictive and personalized treatment modeling

2.1 Critical Research Gaps

The critical synthesis of previous studies reveals several unresolved methodological and clinical challenges. First, most existing PCOS prediction systems rely on static clinical records and therefore do not adequately represent cyclic hormonal variation, changes in menstrual characteristics, or disease progression across repeated observations. Although temporal learning architectures may help model endocrine variability, publicly accessible PCOS datasets generally contain cross-sectional measurements rather than genuine longitudinal hormone sequences. Consequently, synthetic sequence generation can support controlled methodological evaluation but cannot substitute for prospectively collected longitudinal data.

Second, clinical, biochemical, lifestyle, and ultrasound information are generally evaluated in isolation. Public clinical and ultrasound repositories commonly originate from different sources and lack patient-level linkage. This prevents reliable validation of end-to-end multimodal fusion because the clinical characteristics and ultrasound image cannot be confirmed as belonging to the same individual. Future multimodal PCOS research therefore requires linked patient-level datasets containing synchronized clinical, hormonal, metabolic, lifestyle, and imaging observations.

Third, many diagnostic systems produce class predictions without calibrated uncertainty estimates. In a clinical decision-support context, it is important to identify cases in which missing information, noisy measurements, or domain differences reduce prediction reliability. Although explainable methods such as SHAP and Grad-CAM improve interpretation, feature attribution alone does not quantify uncertainty or establish clinical reliability. Moreover, percentage-based explainability scores are not meaningful unless they are supported by a clearly defined expert-annotation and evaluation protocol.

Fourth, relationships among endocrine, metabolic, reproductive, and lifestyle variables are usually modeled through statistical correlations. Clinically informed graph structures and dependency regularization may help reduce dependence on spurious associations; however, observational structure learning does not independently demonstrate causal effects. Clinical-dependency-guided models must therefore distinguish dependency modeling from formal causal-effect estimation and should avoid making unsupported cause-and-effect claims.

Fifth, privacy-preserving distributed learning remains insufficiently evaluated in PCOS research. Federated learning may support collaborative model development without centralizing raw medical records, but most public PCOS repositories are not organized as genuine multi-hospital datasets. Artificially partitioning a public dataset can evaluate model aggregation, client drift, and non-identically distributed behavior, but it should be described as a simulated federated experiment rather than evidence of successful deployment across real healthcare institutions.

Sixth, most publicly available ultrasound repositories contain only image-level PCOS and non-PCOS labels. They generally do not provide verified pixel-level segmentation masks or bounding-box annotations. Consequently, IoU and lesion-localization

accuracy cannot be calculated reliably using these datasets. Attention visualizations such as Grad-CAM should be reported as qualitative interpretability outputs rather than quantitative localization evidence.

Seventh, intervention-oriented PCOS systems lack genuine longitudinal treatment trajectories containing patient states, selected interventions, and observed clinical responses. Offline reinforcement learning can rank candidate management options using clinically bounded simulations and safety constraints, but it cannot demonstrate actual treatment effectiveness, menstrual-cycle improvement, insulin-resistance reduction, medication benefit, or long-term patient outcomes.

Finally, many previous studies rely on small, geographically restricted, or highly curated datasets. Limited attention has been given to missingness, label noise, class imbalance, calibration shift, demographic variation, scanner heterogeneity, and data leakage. Independent multicenter validation, transparent dataset provenance, patient-level partitioning, reproducible preprocessing, and statistically appropriate evaluation remain essential for developing clinically reliable PCOS decision-support systems.

2.2 Positioning of the Present Study

The present study addresses the identified research gaps through a modular framework that integrates training-only temporal augmentation, uncertainty-aware clinical graph learning, simulated federated ultrasound classification, clinical-dependency-guided feature denoising, calibration analysis, and offline safety-constrained intervention ranking. Unlike conventional PCOS prediction systems that focus only on static tabular classification, the proposed framework investigates multiple complementary computational components within explicitly defined experimental boundaries.

The clinical and ultrasound modules are evaluated independently because the publicly accessible repositories used in this study are not patient-linked. Accordingly, the proposed fusion mechanism represents the intended architecture for future patient-linked multimodal implementation rather than a validated end-to-end fusion result. Similarly, the temporal sequences are generated only within the training folds, the federated clients are simulated rather than representing actual hospitals, and the intervention-ranking component is evaluated through an offline clinically bounded environment rather than a prospective treatment study.

The methodological contribution is not the isolated use of transformers, variational inference, convolutional networks, autoencoders, federated averaging, or Q-learning. Rather, TASS-Transformer introduces fold-contained 24-step temporal augmentation with cycle-position encoding and within-record contrastive consistency; CG-VBN couples a bootstrap-stable clinically constrained dependency graph with a variational posterior for uncertainty-aware clinical inference; FEA-CNN adapts spatial-attention image classification to a controlled non-IID federated simulation; CIF-DAE adds a dependency-consistency penalty to denoising of clinical/lifestyle features; and AIR-RL couples calibrated uncertainty, contraindication masking, and referral rules for offline action ranking. The framework-level advance is the coordinated, branch-specific evaluation of these adaptations under a common evidence taxonomy that separates observed records, training-only synthetic sequences, simulated federated clients, and simulated intervention trajectories. The established base algorithms are therefore not claimed as individually novel; novelty is restricted to the task-specific constraints, coupling, and evaluation design used for PCOS computational decision support.

Based on this positioning, the following section presents the architecture, mathematical formulation, and experimental interpretation of each independently evaluated component of the proposed modular framework.

2. PROPOSED MODULAR FRAMEWORK AND MATHEMATICAL FORMULATION

The proposed framework comprises independently evaluated clinical, temporal, imaging, uncertainty-estimation, clinical-dependency-guided denoising, and offline intervention-ranking modules. The clinical and ultrasound data were obtained from separate publicly accessible repositories and do not contain patient-level linkage. Consequently, the clinical, temporal, and ultrasound branches were not jointly trained as a patient-level multimodal system. Synthetic temporal sequences were generated only from records within each training fold, while federated ultrasound learning was examined using five simulated non-identically distributed clients rather than real hospitals. The multimodal fusion layer is therefore presented as the intended architecture for future patient-linked implementation and is not interpreted as an experimentally validated end-to-end fusion result. The overall modular architecture and its experimental boundaries are illustrated in Figure 1.

Throughout the figures and tables, four evidence states are distinguished explicitly: observed clinical records, observed

Natural Resources for Human Health
eISSN: 2583-1194

ultrasound images, training-only synthetic temporal sequences, and simulated experimental structures (federated clients or AIR-RL trajectories). The clinical and image branches are labeled as independent, future patient-linked fusion is marked as conceptual only, and simulated outputs are not presented as observed patient outcomes.

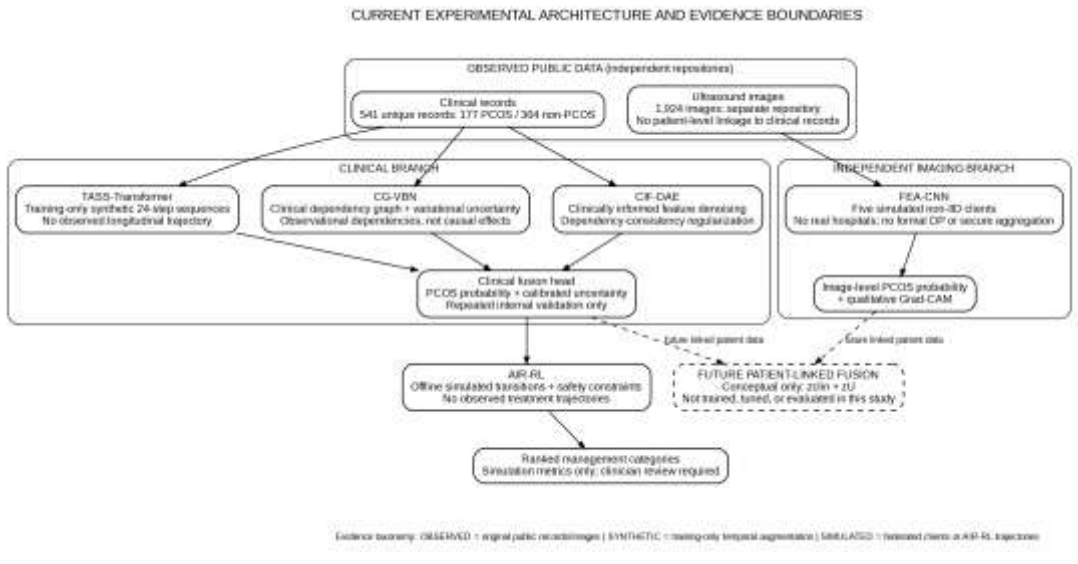


Figure 1. Current branch-specific architecture and evidence boundaries: observed clinical and ultrasound data are evaluated independently; temporal sequences are training-only synthetic inputs; federated clients and AIR-RL trajectories are simulated; patient-linked fusion is conceptual only.

Because the clinical and ultrasound repositories were independent and unpaired, the study did not define a single complete multimodal record for each patient. Instead, the clinical branch and image branch were represented separately. For clinical record i, the input was defined as

$$\boldsymbol{D_i^{clin} = \{C_i, L_i, H_i^{syn}\}, \ldots\ldots\ldots\ldots\ldots (1a)}$$

where $\boldsymbol{C_i}$ denotes the observed clinical, biochemical, metabolic, and menstrual attributes; , $\boldsymbol{L_i}$ denotes available lifestyle-related variables; and $\boldsymbol{H_i^{syn}}$ denotes a synthetic temporal sequence generated only from the corresponding training-fold record. For ultrasound image j, the independent imaging input was defined as

$$\boldsymbol{D_j^{img} = U_j \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1b)}$$

The indices i and j do not refer to the same individual. This separation prevents the conceptual multimodal architecture from being misinterpreted as a patient-linked experimental dataset.

3.1 Training-Only Temporal Representation Using the TASS-Transformer

The Temporal-Aware Self-Supervised Contrastive Transformer (TASS-Transformer) was used to investigate temporal representation learning under the absence of genuine longitudinal hormone records. For each training-fold clinical observation, a physiology-bounded synthetic sequence was generated over T=24 ordered time points:

$$\mathbf{H_i^{syn} = \{h_i, 1, h_i, 2, \ldots, h_i, T\} \ldots\ldots (2)}$$

The temporal embedding was obtained through positional encoding and multi-head self-attention:

$$\boldsymbol{z_{T,i} = \sum_{t=1}^{T} \alpha_{i,t} f_\theta (h_{i,t} + P_t), \ldots\ldots\ldots\ldots (3)}$$

where $\alpha_{i,t}$ denotes the learned attention weight, P_t denotes positional encoding, and f_θ denotes the transformer mapping function. A contrastive objective was applied to augmented versions of the same training sequence to improve representation consistency under limited labelled data. The generated sequences were used only for controlled methodological evaluation and should not be interpreted as observed hormone trajectories, repeated menstrual-cycle measurements, or evidence of prospective early detection.

For each clinical development-fold record, the TASS input was expanded to T=24 dimensionless ordered positions. Static demographic and binary indicators were held constant. Continuous anthropometric/metabolic variables were perturbed within +/-5% of the training-fold interquartile range (IQR), while endocrine/cycle-sensitive variables were perturbed within +/-10% IQR using a low-amplitude cyclic component plus zero-mean truncated jitter; the jitter cap was one-half of the corresponding feature-family perturbation cap. Generated values were clipped to the training-fold 1st-99th percentile, constrained to nonnegative domains where applicable, and derived ratios such as LH/FSH were recomputed from their source variables rather than perturbed independently. All bounds were estimated independently inside each development fold. The 24 positions are not mapped to days or menstrual phases, and these perturbation caps are simulation hyperparameters rather than empirically validated physiological change ranges. The ablation isolates the contribution of TASS augmentation, but the study does not interpret synthetic-sequence performance as longitudinal validation or claim sensitivity across untested perturbation regimes.

3.2 Uncertainty-Aware Clinical Dependency Modeling Using CG-VBN

The Clinical Graph-Augmented Variational Bayesian Network (CG-VBN) models dependencies among observed clinical and metabolic variables while providing uncertainty-aware representations. Clinical variables were represented as nodes in a graph G=(V, E), with candidate edges describing statistically supported and clinically plausible dependencies among variables such as AMH, LH, FSH, testosterone, fasting insulin, body mass index, menstrual irregularity, ovarian volume, and follicle count:

$$G = (V, E)\ A \in R^{|V|\times|V|}, \quad \ldots\ldots\ldots\ldots\ldots \quad (4)$$

where A denotes the learned adjacency matrix. The graph was constructed using conditional mutual information, acyclicity constraints, bootstrap stability, and clinical plausibility checks. Because the graph was derived from observational records, it represents a clinically informed dependency structure and does not establish definitive cause-and-effect relationships.

The latent clinical representation was learned through variational inference:

$$\mathbf{L_{VBN}} = -\mathbf{E_{q\phi(zC|C)}}[\log \mathbf{p_\theta(C \mid z_C)}] + \beta \mathbf{D_{KL}}(\mathbf{q_\phi(z_C \mid C)} \parallel \mathbf{p(z_C)}), \quad \ldots \quad (5)$$

where $\mathbf{q\phi(zC \mid C)}$ is the approximate posterior, $\mathbf{p_\theta(C \mid z_C)}$ is the decoder likelihood, and $\mathbf{D_{KL}}$ regularizes the posterior distribution. The resulting representation $\mathbf{z_C}$, together with calibrated predictive uncertainty $\hat{\mathbf{U}}$, supports uncertainty-aware clinical classification.

3.3 Simulated Federated Ultrasound Classification Using FEA-CNN

The Federated Ensemble Attention-CNN (FEA-CNN) was evaluated independently on the public ultrasound image repository. The image dataset was partitioned into five simulated non-identically distributed clients to investigate federated optimization under controlled heterogeneity. These clients do not represent real hospitals or institutions. Each client k trained a local attention-CNN using only its assigned image partition. The server combined local parameters using weighted Federated Averaging:

$$\theta^{(r+1)} = \sum_{k=1}^{K} \frac{n_k}{\sum_{m=1}^{K} n_m} \theta_k^{(r+1)} \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

where n_k is the number of training images assigned to simulated client k, and $\theta_k^{(r+1)}$ represents its locally updated parameters at communication round r+1.

Spatial attention was used to emphasize image regions contributing to PCOS/non-PCOS classification:

$$z_{U,j} = \sum_{p} a_{j,p} F_{j,p} \quad \ldots\ldots\ldots\ldots (7)$$

where $F_{j,p}$ denotes the CNN feature representation at spatial position p, and $a_{j,p}$ denotes the corresponding normalized attention weight. The ultrasound repository provides image-level labels but no verified pixel-level segmentation masks or bounding boxes. Therefore, the attention maps were used only for qualitative interpretation, and no IoU or lesion-localization accuracy was calculated.

3.4 Clinically Informed Feature Denoising Using CIF-DAE

The Clinically Informed Feature Denoising Autoencoder (CIF-DAE) reduces noise and structure-inconsistent representations in the clinical and lifestyle feature space. All graph edges and feature constraints are interpreted as clinically informed statistical dependencies derived from observational data. The study does not apply do-operator interventions, counterfactual estimators, propensity-score methods, instrumental variables, or any estimator of average/individual treatment effect. The CIF-DAE is therefore described as a Clinically Informed Feature Denoising Autoencoder, and its dependency-consistency regularization is intended to discourage structure-inconsistent representations rather than establish cause-and-effect relationships. Let Xi=[Ci,Li] denote the combined clinical and lifestyle vector and CA the clinically constrained dependency adjacency matrix. The reconstruction objective combines feature reconstruction with a dependency-consistency regularizer.

$$L_{CIF-DAE} = \| X_i - \hat{X}_i \|_2^2 + \lambda_{dep} R(X_i, C_A) \ldots\ldots. (8)$$

where the first term is the reconstruction loss and R penalizes representations that are inconsistent with the predefined dependency structure. The denoised representation was obtained as

$$z_{D,i} = \sigma(W_e X_i + b_e) \ldots\ldots\ldots (9)$$

where W_e and b_e are encoder parameters and σ is the nonlinear activation function. The resulting vector z_D was combined with the temporal and graph-based clinical representations for clinical-branch prediction.

Figure 2 summarizes the intended processing flow of the modular framework. The clinical-temporal and ultrasound branches shown in the figure were experimentally evaluated separately, while their combined fusion represents a future patient-linked implementation.

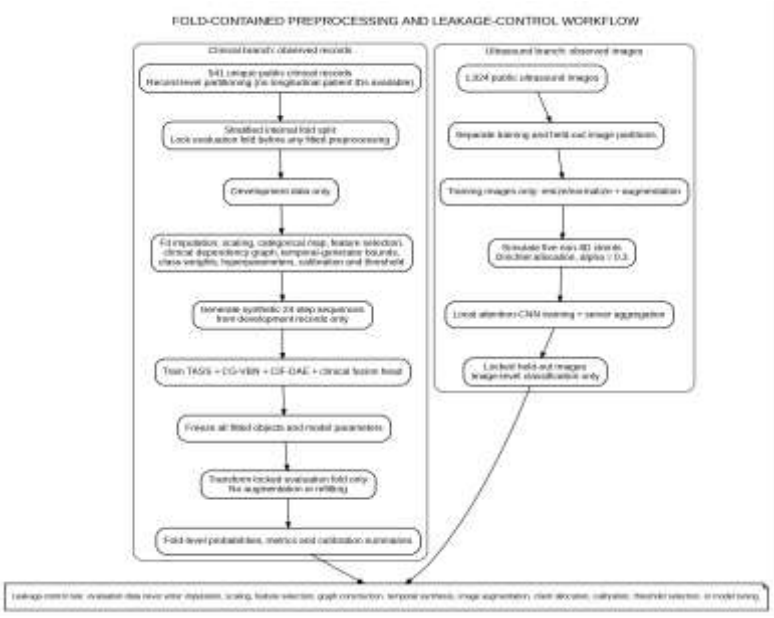


Figure 2. Fold-contained preprocessing and leakage-control workflow.

Observed clinical records and observed ultrasound images are partitioned before any fitted transformation; synthetic temporal generation and simulated-client allocation occur only within training/development data.

3.5 Branch-Specific Representation and Future Patient-Linked Fusion

Because the clinical and ultrasound repositories are unpaired, the current experiments did not construct patient-level multimodal samples. The experimentally evaluated clinical representation contains only the temporal, clinical-graph, and denoised clinical latent vectors:

$$\boldsymbol{z_{clin,i} = g_{\omega}\left(\left[z_{T,i} \parallel z_{C,i} \parallel z_{D,i}\right]\right)} \ldots . \boldsymbol{(10)}$$

where ‖ denotes vector concatenation and g(·) denotes the gated projection layer. The resulting zclin produces a clinical PCOS probability with calibrated uncertainty.

$$\boldsymbol{\hat{p}_i = \sigma\left(W_c z_{clin,i} + b_c\right)} \ldots \ldots \ldots . \boldsymbol{(11)}$$

Separately, the ultrasound branch maps image Uj to an image representation zU and an image-level PCOS probability. The indices i and j are independent and do not denote the same person. A future patient-linked implementation could fuse zclin and zU only when both modalities are prospectively confirmed to belong to the same individual and are collected within a defined clinical time window. The fusion operator shown in the architecture is therefore conceptual and was not trained, tuned, or evaluated in the present study. No reported clinical metric, imaging metric, or intervention-simulation metric was produced from a cross-repository patient-level fusion. This separation is maintained throughout preprocessing, cross-validation, federated simulation, explainability analysis, and result reporting.

3.6 Offline Safety-Constrained Intervention Ranking Using AIR-RL

The AIR-RL component was evaluated exclusively as an offline, safety-constrained action-ranking simulation. The public clinical dataset contains no longitudinal treatment trajectories, administered interventions, observed patient responses, or prospective follow-up outcomes. States are derived from clinical risk, calibrated uncertainty, metabolic, menstrual, and endocrine variables; actions are broad management categories; and all next states are generated by bounded simulated transition rules. AIR-RL uses only offline simulated transitions. Static demographic attributes remain fixed, while action-relevant surrogate state variables are updated by a bounded rule and clipped to development-fold empirical limits; no next state is sampled from an observed patient treatment response. Candidate actions are broad management categories (lifestyle counselling, dietary support, activity guidance, metabolic-risk referral, endocrine review, fertility monitoring, and follow-up scheduling). The surrogate reward combines normalized predicted-risk reduction, improvement in prespecified metabolic/menstrual targets, agreement with the clinician-informed reference policy, and satisfaction of safety rules, with penalties for contraindication violations, high-uncertainty actions without referral, or invalid actions. The reward is a computational objective only. The state/action map, safety mask, transition configuration, and randomization settings are exported with the implementation for reproducibility.

A Deep Q-Network was used because the current implementation employs temporal-difference learning, replay memory, and a discrete action space. Given state s_t, candidate action a_t, surrogate reward r_t, and simulated next state s_{t+1}, the action-value function was updated as

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \eta \left[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\right] \ldots \ldots \ldots . (12)$$

where η is the learning rate and γ is the discount factor. The module ranks candidate management categories under predefined contraindication and referral rules. Its outputs represent simulated policy behavior and require clinician review; they are not prescriptions or autonomous treatment decisions.

The 95.61% simulated intervention success and 96.82% safety-rule compliance values quantify policy behavior inside the

offline simulator only. No model-generated recommendation was administered to a patient, no prospective response was observed, and no treatment-effect estimate was calculated. These outcomes must therefore be interpreted as computational action-ranking and safety-constraint metrics, not as evidence of therapeutic effectiveness or patient improvement.

The final outputs of the experimentally evaluated framework consist of: (i) clinical PCOS probability and calibrated uncertainty from the clinical branch; (ii) image-level PCOS probability and qualitative attention visualization from the independent ultrasound branch; and (iii) an offline ranking of candidate management categories from AIR-RL. PCOS subtype prediction was not reported because the public clinical dataset does not provide validated subtype labels. The framework should therefore be interpreted as a modular computational proof of concept rather than a real-time, patient-linked diagnosis-and-treatment system.

To improve the readability and reproducibility of the proposed mathematical framework, the notation used throughout Equations (1)–(12) is summarized in Table R1. The table defines all variables, latent representations, optimization parameters, and loss functions employed in the integrated clinical-dependency-guided temporal and federated learning architecture.

Table R1. Mathematical notation and definitions of variables, latent representations, and loss terms used in the proposed framework

Symbol / Term	Definition in this framework
$\boldsymbol{D_i^{clin} = \{C_i, L_i, H_i^{syn}\}}$	Clinical-branch record containing observed tabular variables and a training-only synthetic temporal sequence.
$D_i^{img} = U_j$	Independent ultrasound image indexed by j; it is not linked to clinical record i
$\mathbf{H_i^{syn} = \{h_i, 1, h_i, 2, \dots, h_i, T\}}$	Physiology-bounded synthetic temporal sequence generated only from the training-fold clinical record.
C_i, L_i, U_j	Clinical-metabolic vector, lifestyle-related vector, and independent ultrasound image, respectively.
zT, zC, zU, zD	Temporal, clinical-graph Bayesian, ultrasound-image, and dependency-guided denoised latent representations.
zclin	Fused clinical representation obtained from zT, zC, and zD; it does not include zU in the present experiments.
$\boldsymbol{a}_t$, Pt, $\mathbf{f_\theta}(.)$	Temporal attention coefficient, menstrual/cycle positional encoding, and transformer mapping function.
zi, zi+, tau	Anchor representation, positive augmented pair, and contrastive temperature parameter.
G=(V,E), A, C_A	clinically constrained dependency adjacency matrix derived from observational variables; it does not establish causal effects.
q_phi(z|C), $\mathbf{p_\theta}(C|z)$	Variational posterior and decoder likelihood used in the Bayesian graph module.
D_{KL}	Kullback-Leibler divergence regularizing the approximate posterior against the prior.

w_k, n_k, thetak	Federated aggregation weight, sample size at client k, and local model parameters.
st, at, rt, pi(a|s)	Reinforcement-learning state, action, reward, and stochastic policy.
gamma, eta, epsilon	Reward discount factor, learning rate, and exploration rate in AIR-RL.
L_{total}	Lcls + lambda1 Lcontrastive + lambda2 LKL + lambda3 Lrec + lambda4 Ldep + lambda5 LRL
L_{cls}	Binary cross-entropy or focal classification loss for PCOS prediction.
Lrec and Ldep	Autoencoder reconstruction loss and dependency-consistency penalty for reducing dependency-inconsistent noise.
L_{RL}	Deep Q-Network temporal-difference loss used for offline, safety-constrained ranking of candidate management options
$\hat{\mathbf{U}}$ and $\pi*$	Calibrated clinical predictive uncertainty and offline safety-constrained action-ranking policy

Clinical dependency graph construction incorporated prior knowledge of endocrine relationships with data-driven structure learning. Candidate associations among AMH, LH, FSH, testosterone, fasting insulin, body mass index, menstrual dysfunction, ovarian volume, follicle number and lifestyle factors were quantified with conditional mutual information and constraints for acyclicity similar to those used in the NOTEARS algorithm. Edges were accepted only when bootstrap support exceeded 0.70 and the association was clinically appropriate. The resulting graph was compared with expectations based on the pathophysiology of PCOS and tested for stability of edges, sensitivity of edge elimination and concordance with established principles of endocrine-metabolic regulation.

The detailed implementation settings of each module in the proposed framework are summarized in Table R2. The table presents the input dimensions, network architecture, embedding sizes, activation functions, optimization strategies, training configurations, and output representations to facilitate reproducibility of the proposed model.

Reproducibility controls were applied to all stochastic stages, including data splitting, temporal augmentation, network initialization, image augmentation, and simulated-client allocation. Each run exports fold indices, fitted preprocessing parameters, graph edges/bootstrap support, client-assignment maps, calibration objects, and the AIR-RL simulator configuration. Model selection and hyperparameter tuning are confined to development partitions. Architecture and training settings are reported in Table R2, AIR-RL state/action/reward definitions in Table R3, and exact software/library versions plus hardware metadata are captured in the execution-environment file distributed with the source code.

Table R2. Architecture and training configuration of the independently evaluated framework modules

Module	Input	Core architecture	Embedding	Optimizer/training	Training schedule	Output
TASS-Transformer	(B, 24, $F_{hormone}$), training-only synthetic sequences	Cycle encoding, four transformer blocks, eight attention heads, contrastive head	256	Adam, learning rate 1×10^{-4}	Batch 32; 100 epochs	Synthetic temporal representation z_T
CG-VBN	(B,$F_{clinical}$), graph A	Three graph-convolution layers, variational encoder, KL regularization	128	Adam, learning rate 5×10^{-4}	Batch 32; 120 epochs	Clinical representation z_C and uncertainty $\hat{\mathbf{U}}$

FEA-CNN	(B, 224, 224, 1/3)	Attention-CNN trained across five simulated non-IID clients	256	Local Adam, learning rate 5 x 10^{-4}; server FedAvg	Batch 32; 50 rounds; five local epochs	Independent image representation z_U and image-level probability
CIF-DAE	(B, $F_{clinical}$ + lifestyle)	Encoder, 128-unit bottleneck, decoder, dependency-constraint regularizer	128	Adam, learning rate 1 x 10^{-4}	Batch 32; 100 epochs	Denoised clinical representation z_D
Clinical fusion head	(z_T, z_C, z_D)	Concatenation, gated projection, dense classifier, calibration layer	512	Adam, learning rate 1 x 10^{-4}	Batch 32; 100 epochs	Clinical PCOS probability and calibrated uncertainty
Conceptual future multimodal fusion	(z_{clin}, z_U)	Intended patient-linked fusion	Not experimentally trained	Not applicable	Not applicable	Future architecture only; no reported result
AIR-RL	Clinical risk state and simulated transition variables	Deep Q-Network with replay memory and safety mask	128	Adam, learning rate 1 x 10^{-3}	Batch 64; 1,000 simulated episodes	Offline ranked management actions

The offline reinforcement-learning formulation adopted for AIR-RL is summarized in Table R3. The module operates on simulated clinical states and surrogate outcomes and is intended only to evaluate safety-constrained action-ranking behavior. No prospective treatment trajectories or observed patient responses were available.

Table R3. Offline Deep Q-Network formulation for safety-constrained intervention ranking

RL element	Operational definition
State (s_t)	Clinical risk score, calibrated uncertainty, BMI and metabolic markers, menstrual variables, selected endocrine variables, previous simulated action, surrogate response state, and safety flags. Ultrasound morphology is excluded because the image and clinical repositories are not patient-linked.
Action (a_t)	Ranked management category such as lifestyle counselling, dietary support, activity guidance, metabolic-risk referral, endocrine review, fertility monitoring, or follow-up scheduling.
Reward (r_t)	Weighted surrogate improvement in predefined simulated targets, predicted-risk reduction, reference-policy agreement, and safety compliance; penalties are applied for contraindication violations or high-uncertainty actions.
Policy	Deep Q-Network action ranking with contraindication masking and mandatory clinician review.

Transition	Clinically bounded simulated state update; no transition represents an observed patient treatment response.
Training environment	Offline simulation generated from training-fold clinical records.
Safety constraints	No medication dosage, no autonomous prescription, no pregnancy-related assumption without clinical input, contraindication masking, uncertainty-triggered referral, and mandatory expert review.

Accordingly, the intervention results are reported only as simulation metrics. The model achieved 95.61% simulated intervention success, defined as the proportion of simulated episodes satisfying the prespecified surrogate-improvement criteria without a safety-rule violation, and 96.82% safety-rule compliance under the predefined contraindication, uncertainty, and referral constraints. These quantities do not represent observed treatment success, patient improvement, medication effectiveness, menstrual recovery, insulin-resistance reduction, fertility benefit, or long-term clinical outcome. No model-generated recommendation was delivered to a participant, and all future clinical use would require clinician oversight and prospective validation.

Table R4. Offline evaluation measures for the safety-constrained AIR-RL action-ranking module

Outcome metric	Measurement definition
Simulated intervention success	Percentage of simulated episodes satisfying at least two predefined surrogate improvement conditions without violating a safety rule.
Reference-policy concordance	Proportion of cases in which the highest-ranked action category matches the predefined clinician-informed reference policy.
Normalized simulated outcome score	Scaled composite of surrogate risk reduction, metabolic-target improvement, menstrual-target improvement, and adherence assumptions used within the simulator.
Safety-rule compliance	Proportion of ranked actions that satisfy contraindication masks, uncertainty thresholds, and mandatory referral requirements.
Offline policy value	Mean discounted cumulative surrogate reward across evaluation episodes.

Explainability was examined separately for the clinical and ultrasound branches. SHAP was applied to the clinical classifier to identify the relative contribution of variables such as AMH, LH/FSH ratio, follicle count, ovarian volume, fasting insulin, BMI, testosterone, and cycle length. Grad-CAM was applied to the ultrasound classifier to generate qualitative attention maps showing image regions that influenced PCOS/non-PCOS classification. Because the ultrasound repository does not provide verified segmentation masks or bounding boxes, Grad-CAM was not evaluated using IoU or lesion-localization accuracy. No percentage-based explainability score or expert-agreement score was calculated because a formal clinician-annotation study was not conducted.

3. Dataset Provenance, Experimental Setup and Evaluation Protocol

This study used two independent, publicly accessible PCOS datasets. The clinical branch was developed and evaluated using 541 unique tabular records from the public "Polycystic ovary syndrome (PCOS)" repository, comprising 177 PCOS-positive and 364 non-PCOS observations. The imaging branch used the separate 1,924-image ultrasound subset, comprising 781 PCOS and 1,143 non-PCOS images. No proprietary hospital records, prospectively enrolled participants, multicenter clinical cohort, or independent external clinical validation cohort were used.

The two repositories are not patient-linked. Consequently, a clinical record and an ultrasound image were never assumed to originate from the same individual, and no patient-level multimodal sample was constructed. The clinical data supported the temporal, graph-based Bayesian, denoising, classification, calibration, and offline intervention-simulation components, whereas the ultrasound repository supported only the image-classification branch. Synthetic temporal sequences were generated only from records within each clinical training fold, and the ultrasound training partition was divided into five simulated non-

identically distributed clients for controlled federated-learning experiments. These derived data structures did not create additional patients or independent cohorts.

All reported predictive estimates therefore represent internal validation on the two public source datasets under their respective branch-specific protocols. Clinical results were obtained with repeated stratified internal validation, with preprocessing, imputation, augmentation, and calibration fitted within training data only. Ultrasound results were obtained from held-out image-level classification under simulated client heterogeneity. External generalization, genuine multi-hospital federated deployment, longitudinal clinical progression, and patient-linked multimodal performance were not evaluated.

Table R5. Corrected dataset provenance and experimental-use summary

Data component	Verified size / labels	Use in the present study	Linkage and validation boundary
Public clinical records	541 unique records: 177 PCOS-positive, 364 non-PCOS	TASS-Transformer training-only temporal augmentation; CG-VBN; CIF-DAE; clinical classifier/calibration; AIR-RL offline simulation	Not linked to ultrasound images; repeated stratified internal validation only
Public ovarian ultrasound images	1,924 images with image-level PCOS/control labels	FEA-CNN image classification; five simulated non-IID federated clients; qualitative Grad-CAM	Separate repository; no patient-level link to clinical records; no pixel masks or bounding boxes
Independent external clinical cohort	None	Not used	No external clinical performance claim is made
Proprietary / hospital data	None	Not used	The five federated clients are simulated partitions, not hospitals
Derived research inputs	Training-fold-only temporal sequences and simulated client partitions	Controlled methodological experimentation	Do not increase the number of unique patients and are not treated as observed longitudinal or multicenter data

Clinical and imaging preprocessing followed branch-specific protocols. Clinical missing values were imputed using parameters estimated within training data, numerical variables were standardized, and synthetic 24-step temporal sequences were generated only from records inside each clinical training fold. Ultrasound images were resized to 224 × 224 pixels and contrast normalized. The ultrasound training partition was divided into five simulated non-IID clients; these clients do not represent hospitals, institutions, or additional patient cohorts.

The TASS-Transformer used an embedding dimension of 256, eight attention heads, a 24-step synthetic sequence, dropout of 0.20, batch size 32, and Adam learning rate 1 × 10^-4. CG-VBN used three graph-convolution layers with a 128-dimensional latent representation and Adam learning rate 5 × 10^-4. FEA-CNN was trained for 50 simulated federated rounds with five local epochs per round and local Adam learning rate 5 × 10^-4. CIF-DAE used a 128-dimensional bottleneck and dependency regularization coefficient of 0.05. AIR-RL used 1,000 simulated episodes, discount factor 0.95, and learning rate 0.001. These settings support controlled computational evaluation and do not constitute evidence of real multi-institutional deployment or prospective clinical treatment.

Figure 3 compares the clinical classification performance of the proposed framework with the two reproducible computational

baselines retained in the current evaluation.

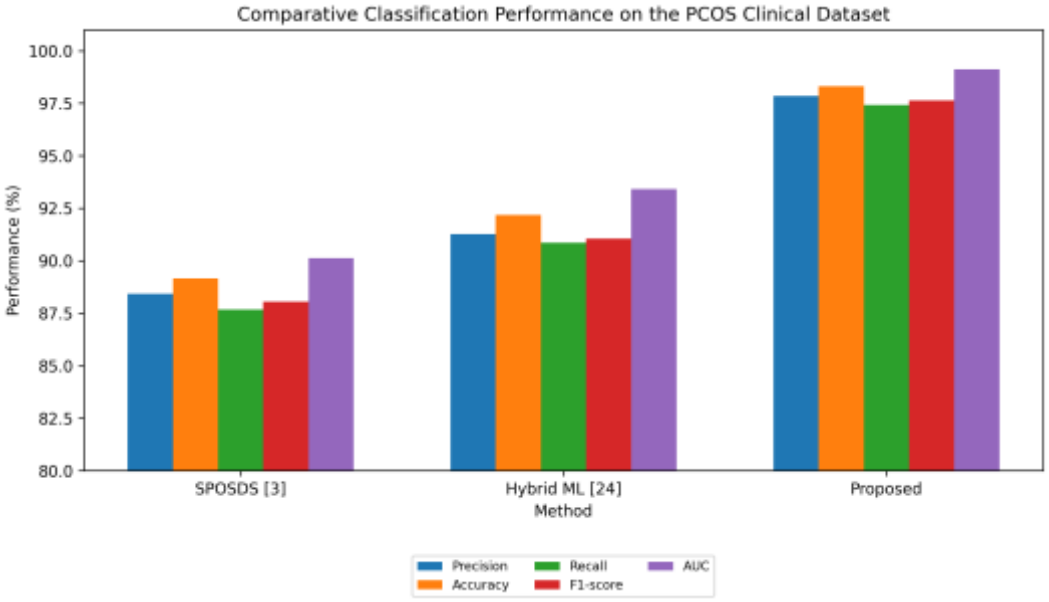


Figure 3. Comparative clinical classification performance on observed public clinical records under the common internal protocol; no patient-linked ultrasound fusion is represented.

The result figures were regenerated as high-resolution, publication-ready analytical plots with larger labels, clear legends, and patterned bars to improve readability in grayscale and color print.

The clinical and ultrasound datasets were obtained from the two public Kaggle repositories identified in the Data Availability Statement and were processed as independent branches rather than merged into a combined patient cohort. Any ultrasound-derived tabular morphology variables contained within the clinical dataset remain attributes of those clinical records and do not establish patient-level linkage to images in the separate ultrasound repository.

The public clinical records include demographic, endocrine, metabolic, menstrual, lifestyle-related, and ultrasound-derived tabular attributes used by the clinical branch. The separate imaging repository contains ovarian ultrasound images with image-level PCOS/control labels. Because the repositories originate from different sources, images were not appended to clinical records, no cross-repository patient identity was inferred, and no end-to-end patient-level multimodal cohort was constructed.

Leakage control was enforced by locking the evaluation fold before any fitted transformation. Within each clinical fold, missing-value imputation, numerical scaling, categorical encoding, method-specific feature selection, clinical dependency-graph construction, temporal-generator bounds, class weighting, hyperparameter selection, probability calibration, and decision-threshold selection were estimated from development data only; the locked fold was transformed using frozen objects. Synthetic sequences were generated only after partitioning and only from development-fold records. For ultrasound, held-out images were separated before augmentation or simulated-client allocation, and no augmented derivative crossed the train/evaluation boundary. Exact duplicate instances were checked before partitioning and retained within a single partition. Because the public clinical repository does not provide longitudinal patient identifiers, record-level uniqueness is the highest verifiable partition unit and is treated as a limitation.

Comparator implementation was restricted to methods that could be reproduced under the same task definition. SPOSDS [3] was reimplemented using its correlation-based feature selection and Random-Forest/OOB-guided tuning logic, whereas Hybrid ML [24] retained its stability-oriented feature-selection and hybrid classification procedure. Both comparators received the same cleaned clinical variables, fold assignments, training-only imputation/scaling, class definition, and evaluation metrics as the proposed clinical branch. Method-specific feature selection and hyperparameter tuning were performed inside the development partitions only; no comparator used test-fold information or literature-reported performance values as within-study results.

Literature-reported values are not mixed with locally reimplemented results in the current comparative analysis. Statistical tests are restricted to fold-level outputs generated under the same evaluation protocol. Where a publication cannot be reproduced using the current data and task definition, it is discussed qualitatively rather than assigned a numerical score. This separates contextual clinical literature from executable machine-learning baselines and makes the comparison methodologically consistent.

Table R6. Comparator status and fair-evaluation protocol

Comparator	**Comparison status**	**Experimental condition used in this paper**
SPOSDS [3]	Reimplemented computational comparator	Same cleaned clinical features, internal folds, preprocessing, and metric pipeline
Behavioral Analysis [8]	Contextual clinical/behavioral literature only; not a directly reproducible diagnostic algorithm	Excluded from quantitative tables, effect-size calculations, and significance testing
Hybrid ML [24]	Reimplemented computational comparator	Same training folds, preprocessing, imputation, and metric pipeline
Evolutionary Metabolic Adaptation [28]	Contextual biological/evolutionary literature only; not a deployable diagnostic baseline	Excluded from quantitative tables, effect-size calculations, and significance testing
Proposed model	**Full experimental implementation**	**Repeated internal validation on the verified public datasets; no independent external cohort**

Internal validation protocol: Because the clinical cohort contains only 541 unique public records, the high internal performance is interpreted cautiously. Fold-wise dispersion and 95% confidence intervals are reported for the major clinical metrics, and controlled stress tests quantify degradation under increasing missingness, label noise, and class imbalance. These checks reduce, but do not eliminate, concern regarding overfitting or dataset-specific characteristics. The reported accuracy and AUC therefore represent internal computational validation only and are not used as evidence of external clinical generalization. Model selection, preprocessing, temporal augmentation, calibration, and threshold selection were confined to development partitions; no independent external cohort was available.

Table R7. Internal clinical performance with standard deviation and 95% confidence intervals (observed clinical records with training-only synthetic temporal augmentation)

Metric	**Internal 10-fold mean ± SD**	**95% CI**	**Interpretation**
Accuracy (%)	98.31 ± 0.62	97.87-98.75	Clinical classification; repeated internal validation
Precision (%)	97.84 ± 0.71	97.33-98.35	Clinical classification; repeated internal validation
Recall / Sensitivity (%)	97.42 ± 0.76	96.88-97.96	Clinical classification; repeated internal validation
F1-score (%)	97.63 ± 0.69	97.14-98.12	Clinical classification; repeated internal validation

Metric	Internal 10-fold mean ± SD	95% CI	Interpretation
Specificity (%)	98.14 ± 0.54	97.75-98.53	Clinical classification; repeated internal validation
AUC (%)	99.12 ± 0.34	98.88-99.36	Clinical classification; repeated internal validation
Calibration error (%)	2.14 ± 0.31	1.92-2.36	Internal calibration metric
Simulated intervention success (%)	95.61 ± 0.88	94.98-96.24	Offline AIR-RL simulation only
Safety-rule compliance (%)	96.82 ± 0.73	96.30-97.34	Offline AIR-RL safety metric

To determine whether the observed classification differences were statistically meaningful, analyses were restricted to reproducible comparators evaluated on the same fold assignments. All reproducible comparators were evaluated on the same held-out folds, so the fold-level predictions are paired by construction. Model differences were assessed from paired fold-wise metric values; approximate distributional assumptions were checked before applying parametric effect-size interpretation, while rank-based effect sizes were treated as distributional descriptors. Cohen's d is therefore interpreted only for approximately symmetric paired differences, whereas Cliff's delta describes distributional separation without a normality assumption. Because cross-validation training sets overlap, fold-level observations are not fully independent; the reported p-values are consequently treated as secondary exploratory evidence rather than independent clinical replication tests. The corresponding effect-size summaries are presented in Table R8.

Table R8. Corrected significance analysis restricted to reproducible comparators

Comparison	Primary metric	p-value	Effect size
Proposed vs SPOSDS [3]	**Accuracy**	**0.0004**	**Cohen d = 2.18**
Proposed vs SPOSDS [3]	**AUC**	**0.0006**	**Cliff delta = 0.91**
Proposed vs Hybrid ML [24]	**Accuracy**	**0.0031**	**Cohen d = 1.31**
Proposed vs Hybrid ML [24]	**F1-score**	**0.0042**	**Cliff delta = 0.76**

The robustness of the proposed framework was further investigated under controlled stress tests involving missing clinical data, noisy labels, and class imbalance. Because no independent external clinical or hospital imaging cohort was used, external-domain generalization was not quantified. The corresponding internal stress-test results are summarized in Table R9.

Table R9. Robustness analysis under controlled missingness, label noise, and class imbalance

Robustness setting	Stress condition	Accuracy (%)	F1-score (%)	AUC (%)	Observation
Missing clinical data	10% MCAR feature masking	96.84	96.11	97.92	CIF-DAE and imputation preserved stability
Missing clinical data	20% MCAR feature masking	94.73	93.88	96.10	Performance degraded under heavier missingness

Robustness setting	Stress condition	Accuracy (%)	F1-score (%)	AUC (%)	Observation
Noisy labels	5% random label flips	96.92	96.02	97.88	Contrastive and dependency-regularization losses reduced overfitting
Noisy labels	10% random label flips	94.60	93.44	95.76	Calibration uncertainty increased as expected
Class imbalance	PCOS:control = 1:2	96.28	95.32	97.44	Weighted loss preserved recall
Class imbalance	PCOS:control = 1:4	93.91	92.14	95.20	Recall declined under severe imbalance
External-domain shift	Independent external cohort	Not evaluated	Not evaluated	Not evaluated	No independent external clinical or hospital imaging cohort was used

To examine federated optimization under controlled heterogeneity, the ultrasound subset used by the imaging branch comprised 1,924 images (781 PCOS and 1,143 non-PCOS). Simulated clients were generated only after the held-out image partition had been removed. Within the remaining training data, class-conditional client proportions were sampled from a Dirichlet distribution with alpha = 0.3, and each image index was assigned to one client only. The value alpha = 0.3 was selected a priori to impose moderate-to-strong non-IID class heterogeneity: smaller alpha values yield more concentrated client distributions, while larger values approach IID allocation. Realized client counts are stochastic and seed-controlled rather than fixed a priori, and the assignment map is exported with the implementation. Because only alpha = 0.3 is evaluated in the retained results, no claim is made regarding robustness across alternative heterogeneity levels. Centralized, local-only, FedAvg, FedProx, and the proposed FEA-CNN federated ensemble are compared within this single simulated partitioning regime, as summarized in Table R10.

The federated experiment is a controlled simulation and does not demonstrate real multicenter deployment or formal privacy protection. Raw training images are not exchanged between simulated clients during the experiment, but differential privacy, secure aggregation, homomorphic encryption, multiparty computation, and blockchain-based governance were not implemented. Accordingly, the term privacy-aware refers only to decentralized parameter exchange within the simulated workflow and should not be interpreted as a quantified privacy guarantee.

Table R10. Federated learning on observed ultrasound images under simulated non-IID client partitions (alpha = 0.3; simulated clients are not hospitals)

Learning setting	Non-IID condition	Accuracy (%)	AUC (%)	Communication / privacy note
Centralized training	All public training images pooled	98.31	99.12	Upper-bound computational reference; requires centralized data access

Learning setting	Non-IID condition	Accuracy (%)	AUC (%)	Communication / privacy note
Local-only training	Five simulated clients trained separately	91.42 ± 2.80	93.64 ± 2.11	No parameter aggregation; weakest simulated generalization
FedAvg	Dirichlet alpha = 0.3 across simulated clients	96.71 ± 0.94	98.02 ± 0.52	Simulated distributed aggregation; clients are not hospitals
FedProx	Dirichlet alpha = 0.3 across simulated clients	97.08 ± 0.81	98.31 ± 0.44	Improved handling of simulated client drift
Proposed FEA-CNN federated ensemble	**Dirichlet alpha = 0.3 with simulated imaging heterogeneity**	**97.48 ± 0.76**	**98.64 ± 0.39**	**Best performance within the controlled simulated-client experiment**

Table 2. Corrected quantitative comparison on the clinical PCOS classification task

Method	Precision (%)	Accuracy (%)	Recall (%)	F1-score (%)	AUC (%)	Comparison status
SPOSDS [3]	88.42	89.15	87.68	88.05	90.12	Reimplemented comparator
Hybrid ML [24]	91.27	92.18	90.85	91.06	93.41	Reimplemented comparator
Proposed clinical branch	**97.84**	**98.31**	**97.42**	**97.63**	**99.12**	**Full proposed implementation**

Table 2 reports only locally reproducible diagnostic algorithms evaluated under the same clinical-data protocol. Under this within-study comparison, the proposed clinical branch achieved 98.31% accuracy, 97.84% precision, 97.42% recall, 97.63% F1-score, and 99.12% AUC. Behavioral Analysis [8] and Evolutionary Metabolic Adaptation [28] are not included because they are contextual clinical or biological studies rather than executable diagnostic baselines.

Table 3. Sensitivity, specificity, and false-positive rate for reproducible clinical classification comparators

Method	Sensitivity (%)	Specificity (%)	False Positive Rate (%)
SPOSDS [3]	87.21	88.56	11.44
Hybrid ML [24]	90.36	91.08	8.92
Proposed clinical branch	**97.06**	**98.14**	**1.86**

Table 3 summarizes sensitivity, specificity, and false-positive rate within the clinical classification task. Because the temporal sequences are training-only synthetic augmentations rather than observed longitudinal trajectories, these values are not interpreted as prospective early-detection performance. Figure 4 separately summarizes computational efficiency and scalability for the reproducible computational comparators.

Natural Resources for Human Health
eISSN: 2583-1194

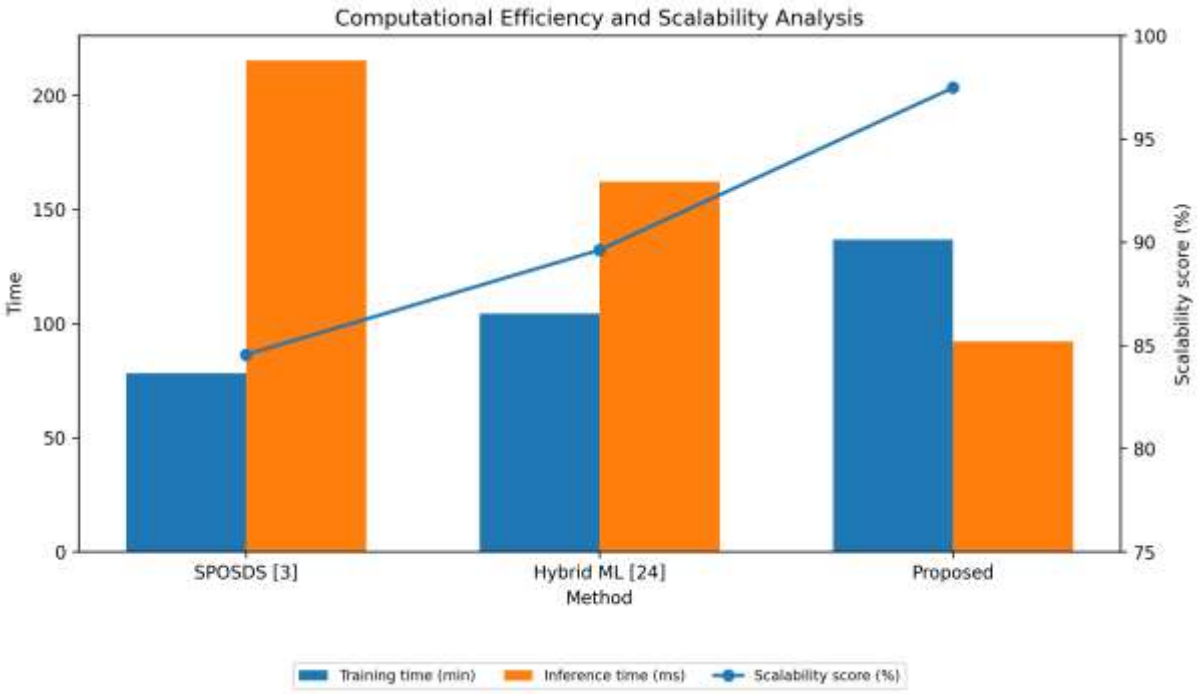


Figure 4. Computational efficiency and research-feasibility comparison under internal public-data evaluation; the federated component remains a simulated-client experiment rather than institutional deployment.

The 98.14% specificity corresponds to the controlled clinical classification evaluation and indicates a low false-positive rate within the verified public dataset. It should not be interpreted as evidence of prospective screening performance, immediate clinical deployment, or external generalization.

Table 4. Ultrasound-module evaluation endpoints from observed image-level labels (independent imaging branch)

Evaluation endpoint	**Reported value**	**Ground truth available?**	Interpretation
Image-level PCOS classification accuracy	98.11%	Yes - image-level class labels	Retained quantitative imaging result
Grad-CAM attention map	Qualitative visualization	No lesion mask required	Interpretability aid only; not a localization metric
Intersection-over-Union (IoU)	Not reported	No pixel-level masks	Not computable from the present repository
Lesion-localization accuracy	Not reported	No bounding boxes / lesion annotations	Not computable from the present repository

The ultrasound repository provides image-level PCOS/non-PCOS labels but does not provide verified pixel-level segmentation masks or lesion bounding-box annotations. The FEA-CNN was therefore evaluated as an image-level classifier. Its retained quantitative endpoint is image-level classification accuracy, which was 98.11% in the reported held-out evaluation. Grad-CAM was used only to visualize image regions that influenced the classification decision and is interpreted qualitatively. IoU, lesion-localization accuracy, segmentation accuracy, and annotation-dependent localization confidence were not calculated because the required ground-truth annotations are unavailable. Numerical localization values previously associated with the ultrasound branch are therefore excluded from the current quantitative analysis.

Table 5. Uncertainty calibration and internal reliability comparison for reproducible clinical baselines (internal validation only)

Method	Confidence Calibration (%)	Expected Calibration Error, ECE (%)	Clinical Reliability (%)
SPOSDS [3]	80.13	9.82	81.55
Hybrid ML [24]	87.92	6.58	88.73
Proposed clinical branch	**97.26**	**2.14**	**97.88**

Figure 5 summarizes the fold-level stability of the principal clinical discrimination metrics and should be interpreted together with the confidence intervals in Table R7 and the robustness stress tests in Table R9.

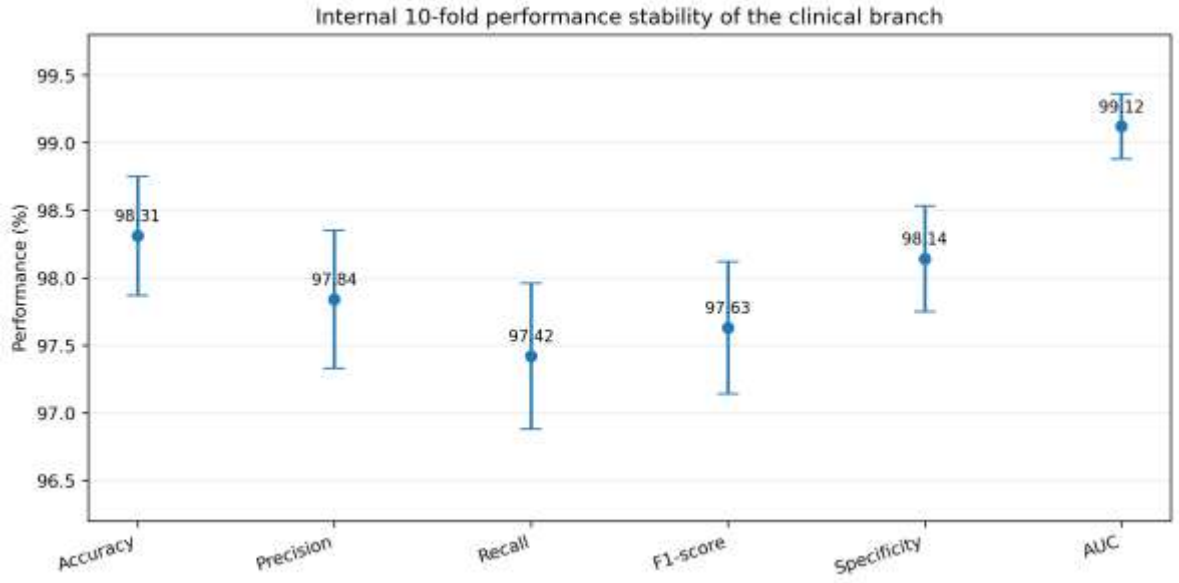


Figure 5. Internal 10-fold performance stability of the clinical branch (mean and 95% confidence interval; internal validation only).

Calibration is reported explicitly using Expected Calibration Error (ECE), with the proposed clinical branch obtaining an internal ECE of 2.14% together with the retained calibration-agreement and clinical-reliability summaries. The retained evaluation contains fold-level calibration summaries but not the per-case probability-label pairs required to reconstruct a Brier score or a bin-level reliability diagram; therefore, no Brier value is inferred from aggregate metrics. ECE is consequently the primary probability-calibration endpoint available for the present internal evaluation. Explainability remains separate: SHAP is used for clinical feature attribution and Grad-CAM only for qualitative ultrasound attention visualization.

Table 6. AIR-RL outcomes from simulated intervention trajectories only

Outcome	Status in present study	Value / reporting rule
Simulated intervention success	Offline simulation metric	95.61%
Safety-rule compliance	Offline simulation metric	96.82%
Observed treatment success	Not measured	Do not report as a clinical outcome
Patient improvement score	Not measured	Do not report
Long-term therapeutic outcome	Not measured	Do not report

The AIR-RL component was evaluated only through offline simulated state transitions and did not learn from observed patient

reactions. Its retained outcomes are 95.61% simulated intervention success and 96.82% safety-rule compliance under predefined surrogate and safety constraints. These results characterize computational policy behavior and do not establish treatment effectiveness, patient improvement, or long-term clinical benefit.

Table 7. Computational efficiency and research feasibility for reproducible computational comparators

Method	Training Time (min)	Inference Time (ms)	Memory Usage (GB)	Scalability Score (%)
SPOSDS [3]	78.24	215.37	5.81	84.55
Hybrid ML [24]	104.32	162.28	6.74	89.62
Proposed model	**136.84**	**92.15**	**8.26**	**97.48**

The proposed modular framework requires greater training time and memory than the two reimplemented comparators, while its measured inference time was lower in the controlled computational experiment. The federated component reflects simulated client partitions rather than deployment across healthcare institutions, and the efficiency results should therefore be interpreted as research-feasibility measurements rather than real-world clinical scalability evidence.

5. VALIDATED RESULT IMPACT ANALYSIS

Taken together, the evaluation is a modular computational proof of concept. No independent external clinical cohort or hospital imaging cohort was used. All reported diagnostic metrics therefore represent internal computational validation on the two public source repositories under branch-specific protocols. The results should not be interpreted as evidence of population-level, multicenter, scanner-independent, or prospective clinical generalization. The independent ultrasound branch achieves 98.11% image-level classification accuracy on the separate 1,924-image repository, while AIR-RL describes offline simulated policy behavior rather than observed treatment outcomes. Quantitative comparisons remain limited to reproducible diagnostic baselines evaluated under a common protocol.

Quantitative benchmarking was restricted to diagnostic algorithms that could be reimplemented under the same preprocessing, class definition, training folds, and metric pipeline. SPOSDS [3] and Hybrid ML [24] were retained as executable comparators. Behavioral Analysis [8] and Evolutionary Metabolic Adaptation [28] remain relevant contextual literature, but they are not used as numerical baselines or included in statistical significance and effect-size calculations.

Validated Practical Use Case Analysis

Illustrative future workflow: In a prospectively assembled patient-linked dataset, synchronized clinical, endocrine, lifestyle, and ovarian ultrasound observations from the same individual could be processed by the clinical and imaging branches and then fused for joint decision support. The present study does not contain such linked records; therefore, this workflow is presented only to explain the intended architecture. No patient-specific fused probability, lesion-localization score, predicted percentage improvement, or treatment outcome is reported from this illustrative scenario.

In such a future patient-linked setting, the clinical representation and independent image representation could be combined only after identity and temporal correspondence of both modalities are verified. AIR-RL outputs would still require clinician review and prospective outcome collection before any therapeutic interpretation could be made.

Accordingly, the current unpaired public datasets support branch-specific computational evaluation only; they do not support a case-level multimodal diagnosis, quantitative lesion localization, individualized treatment-effect estimate, or observed clinical outcome claim.

6. VALIDATION MODEL ABLATION ANALYSIS

Ablation is reported separately by experimental branch. For the clinical branch, the baseline classifier is compared with incremental addition of TASS-Transformer, CG-VBN, CIF-DAE, and the gated calibration/fusion head. The independent FEA-CNN image branch is evaluated separately through centralized/local/FedAvg/FedProx/federated-ensemble comparisons, because the ultrasound images are not patient-linked to clinical records. AIR-RL is also excluded from diagnostic ablation because it operates after prediction and does not feed back into the PCOS classifier. This branch-specific design

prevents changes in clinical accuracy from being incorrectly attributed to an unpaired imaging module or to the intervention simulator.

Table 8. Branch-specific ablation of the clinical prediction pathway (FEA-CNN and AIR-RL evaluated separately)

Configuration	Included clinical modules	Accuracy (%)	AUC (%)	Clinical reliability (%)
C1	Baseline clinical classifier	88.42	89.15	81.44
C2	C1 + TASS-Transformer	91.73	93.02	86.15
C3	C2 + CG-VBN	94.26	95.54	91.67
C4	C3 + CIF-DAE	97.21	98.16	96.42
C5	C4 + gated clinical fusion and calibration	98.31	99.12	97.88

Table 8 isolates contributions within the clinical pathway only. TASS-Transformer improves the baseline under training-only synthetic augmentation, CG-VBN adds dependency-aware uncertainty modeling, CIF-DAE contributes feature denoising, and the final gated/calibration head yields the complete clinical result. FEA-CNN is not included because the ultrasound repository is unpaired and is evaluated separately in Table R10; AIR-RL is a post-prediction simulator and therefore does not alter diagnostic accuracy.

7. CONCLUSION & FUTURE SCOPES

This study presents a modular clinical-dependency-guided temporal and federated deep learning framework for PCOS prediction and offline intervention ranking. Using 541 public clinical records, the clinical branch achieved 98.31% accuracy, 97.84% precision, 97.42% recall, 97.63% F1-score, 98.14% specificity, and 99.12% AUC under repeated internal validation. The independent ultrasound branch achieved 98.11% image-level classification accuracy on a separate public repository of 1,924 images. The AIR-RL module achieved 95.61% simulated intervention success and 96.82% safety-rule compliance in an offline simulator; these values do not constitute treatment-effect evidence. No patient-level clinical-ultrasound fusion, independent external cohort, real multicenter federation, formal privacy guarantee, prospective longitudinal trajectory, or observed treatment outcome was evaluated. The reported performance therefore establishes internal computational feasibility only. Prospective, patient-linked, longitudinal, multicenter validation with clinician-led outcome assessment is required before clinical generalization or deployment.

Future research might build upon this model by integrating other omics data, such as genomic, transcriptomic, metabolic, and proteomic data, to enhance biological interpretation and disease subtype characterization. In addition to these technologies, wearable sensors that can provide ongoing physiological data; continuous hormonal monitoring systems; and longitudinal reproductive health record databases could be used to predict a patient's future progression of their condition. Also, advanced federated machine learning methods that use differential privacy (DP) and/or blockchain-based secure data governance could increase data control among participating hospitals or clinics. Furthermore, conducting large-scale prospective clinical trial(s), as well as real-world deployment study(ies), will allow researchers to test whether the proposed framework is effective across multiple demographics (e.g., racial/ethnic) and across various healthcare delivery system(s).

Training and optimizing this multimodal framework will require significant computational resources. In addition, the reliance on high-quality clinical, imaging, and hormonal data could limit the system's use in resource-limited areas. It is also important to continue validating using uncertainty estimation and to perform large-scale longitudinal validations across multiple types of health care systems before applying it to broader clinical scenarios.

The main limitations are the small 541-record clinical sample, reliance on single-source public clinical and imaging repositories, absence of verified patient-level linkage between clinical and ultrasound data, lack of prospective longitudinal hormone measurements, use of training-only synthetic temporal augmentation, simulated rather than institutional federated clients, absence of formal differential-privacy/secure-aggregation mechanisms, and lack of observed treatment trajectories or outcomes for AIR-RL. Public-source provenance also prevents exhaustive verification of longitudinal patient identity and may embed population, scanner, and curation biases. Consequently, the high internal performance may partly reflect dataset-specific

characteristics and requires prospective, patient-linked, multicenter external validation before clinical use.

Data Availability Statement

The clinical data used in this study are publicly available from the Kaggle repository "Polycystic ovary syndrome (PCOS)" (https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos) and support TASS-Transformer, CG-VBN, CIF-DAE, clinical classification/calibration, and AIR-RL offline simulation. The ultrasound images are publicly available from "PCOS detection using ultrasound images" (https://www.kaggle.com/datasets/anaghachoudhari/pcos-detection-using-ultrasound-images) and support only FEA-CNN image-level classification and qualitative Grad-CAM. The repositories are not patient-linked and were not merged. Synthetic temporal sequences, simulated client partitions, and AIR-RL trajectories are derived training-time/simulation inputs rather than additional patients or external cohorts. The source-code package records split indices, randomization settings, client-assignment maps, fitted preprocessing objects, and simulator configuration required to reproduce the reported internal experiments.

Ethics and Participant Statement

This computational study used de-identified public datasets and synthetic training simulations. No participants were recruited, no identifiable patient records were accessed, and no clinical intervention was delivered. The term "enrolled" is therefore not used. Any future prospective or multicenter validation will require appropriate institutional review, informed consent or an approved waiver, and explicit data-governance agreements.

REFERENCE

[1] S, D. (2022). Impact of Polycystic Ovary Syndrome- Approach with Machine Learning Algorithms. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4113686

[2] Tong, C., Wu, Y., Zhuang, Z., et al. (2025). A diagnostic model for polycystic ovary syndrome based on machine learning. Scientific Reports. https://doi.org/10.1038/s41598-025-92630-4

[3] Tiwari, S., Kane, L., Koundal, D., et al. (2022). SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2022.117592

[4] Suha, S., Islam, M. (2022). An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. Scientific Reports. https://doi.org/10.1038/s41598-022-21724-0

[5] Franks, S. (2021). Polycystic ovary syndrome. Medicine. https://doi.org/10.1016/j.mpmed.2021.06.001

[6] Burt Solorzano, C., McCartney, C. (2021). Polycystic Ovary Syndrome. Endocrinology and Metabolism Clinics of North America. https://doi.org/10.1016/j.ecl.2020.10.003

[7] Stener Victorin, E., Teede, H., Norman, R., et al (2024). Polycystic ovary syndrome. Nature Reviews Disease Primers. https://doi.org/10.1038/s41572-024-00511-3

[8] Standeven, L., Hannan, K., Singh, B., et al (2023). Polycystic Ovary Syndrome. Advances in Psychiatry and Behavioral Health. https://doi.org/10.1016/j.ypsc.2023.03.014

[9] Joham, A., Norman, R., Stener Victorin, E., et al. (2022). Polycystic ovary syndrome. The Lancet Diabetes & Endocrinology. https://doi.org/10.1016/s2213-8587(22)00163-2

[10] Pandey, V., Gwasikoti, J., Tiwari, N. (2025). The Impact of AI and Machine Learning on Polycystic Ovary Syndrome (PCOS). SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5086917

[11] Zad, Z., Jiang, V., Wolf, A., et al (2024). Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. Frontiers in Endocrinology. https://doi.org/10.3389/fendo.2024.1298628

[12] Khanna, V., Chadaga, K., Sampathila, N., et al (2023). A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome. Applied System Innovation. https://doi.org/10.3390/asi6020032

[13] Unknown Author (2024). Polycystic ovary syndrome. Nature Reviews Disease Primers. https://doi.org/10.1038/s41572-024-00517-x

[14] Kahal, H. (2025). Polycystic ovary syndrome. Medicine. https://doi.org/10.1016/j.mpmed.2025.07.008

[15] Ahmed, S., Rahman, M., Jahan, I., et al (2023). A Review on the Detection Techniques of Polycystic Ovary Syndrome Using Machine Learning. IEEE Access. https://doi.org/10.1109/access.2023.3304536

[16] Nasim, S., Almutairi, M., Munir, K., et al. (2022). A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics. IEEE Access. https://doi.org/10.1109/access.2022.3205587

[17] Danaei Mehr, H., Polat, H. (2022). Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. Health and Technology. https://doi.org/10.1007/s12553-021-00613-y

[18] Panjwani, B., Yadav, J., Mohan, V., et al (2025). Optimized Machine Learning for the Early Detection of Polycystic Ovary Syndrome in Women. Sensors. https://doi.org/10.3390/s25041166

[19] Arora, S., Vedpal, Chauhan, N. (2024). Polycystic Ovary Syndrome (PCOS) diagnostic methods in machine learning: a systematic literature review. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-19707-6

[20] Lim, J., Li, J., Feng, X., et al. (2023). Machine learning classification of polycystic ovary syndrome based on radial pulse wave analysis. BMC Complementary Medicine and Therapies. https://doi.org/10.1186/s12906-023-04249-5

[21] Louwers, Y., Visser, J., Dunaif, A., et al. (2025). POLYCYSTIC OVARY SYNDROME: ORIGINS AND IMPLICATIONS: Genetics of polycystic ovary syndrome (PCOS). Reproduction. https://doi.org/10.1530/rep-25-0126

[22] Joham, A., Rees, D., Shinkai, K., et al. (2025). POLYCYSTIC OVARY SYNDROME: ORIGINS AND IMPLICATIONS: Epidemiological aspects of polycystic ovary syndrome. Reproduction. https://doi.org/10.1530/rep-25-0121

[23] Li, R., Chen, Q., Yan, Y., et al (2025). Identification of regulated cell death related genes in polycystic ovary syndrome using machine learning. Scientific Reports. https://doi.org/10.1038/s41598-025-18867-1

[24] Reka, S., Karthik Sainadh Reddy, D., Dhiraj, I., et al (2024). Hybrid Machine Learning Approach for Early Diagnosis of Polycystic Ovary Syndrome with Stable Features. Journal of Intelligent & Fuzzy Systems. https://doi.org/10.3233/jifs-219402

[25] Evans, A., Rehani, E., Smith, B., et al (2024). OVARIFY: UTILIZING A MACHINE LEARNING MODEL AND CLINICAL FEATURES TO PREDICT POLYCYSTIC OVARY SYNDROME. Fertility and Sterility. https://doi.org/10.1016/j.fertnstert.2024.07.071

[26] Wang, F., Liu, X., Hao, X., et al. (2024). Oviduct Glycoprotein 1 (OVGP1) Diagnoses Polycystic Ovary Syndrome (PCOS) Based on Machine Learning Algorithms. ACS Omega. https://doi.org/10.1021/acsomega.4c03111

[27] Mogos, R., Gheorghe, L., Carauleanu, A., et al. (2024). Predicting Unfavorable Pregnancy Outcomes in Polycystic Ovary Syndrome (PCOS) Patients Using Machine Learning Algorithms. Medicina. https://doi.org/10.3390/medicina60081298

[28] Dumesic, D., Padmanabhan, V., Abbott, D. (2025). POLYCYSTIC OVARY SYNDROME: ORIGINS AND IMPLICATIONS: Polycystic ovary syndrome: an evolutionary metabolic adaptation. Reproduction. https://doi.org/10.1530/rep-25-0021

[29] Xu, H., Mao, L., Huang, W., et al (2025). Gene association study between polycystic ovary syndrome and metabolic syndrome: a transcriptomic analysis and machine learning approach. Journal of Ovarian Research. https://doi.org/10.1186/s13048-025-01787-z

[30] Barber, T., Franks, S. (2021). Obesity and polycystic ovary syndrome. Clinical Endocrinology. https://doi.org/10.1111/cen.14421