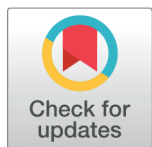


ORIGINAL ARTICLE



OPEN ACCESS

Received: 20/01/2026

Accepted: 16/02/2026

Published: 28/02/2026

Citation: Jagadeesan S, Marannan P (2026) Reliable and High-Precision PCOS Detection Using C2T-Transformer with Uncertainty-Calibrated Learning on Clinical Text Data. Indian Journal of Science and Technology 19(8): 504-524. <https://doi.org/10.17485/IJST/v19i8.79>

* **Corresponding author.**

sathyapriya08@yahoo.com

Funding: None

Competing Interests: None

Copyright: © 2026 Jagadeesan & Marannan. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](https://www.indjst.org/))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Reliable and High-Precision PCOS Detection Using C2T-Transformer with Uncertainty-Calibrated Learning on Clinical Text Data

Sathiyapriya Jagadeesan^{1*}, Praveena Marannan²

¹ Department of Computer Science, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore, Tamil Nadu, India

² Department of Computer Science (Full Stack and Web Development), Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore, Tamil Nadu, India

Abstract

Objectives: To propose a reliable and high-precision deep learning-based PCOS detection model using clinical text data in order to maximize diagnostic accuracy, minimize false predictions and clinician dependency. The model improves robustness under class imbalance and enables scalable, real-time deployment as a support system across diverse healthcare environments. **Methods:** A Causal-Contrastive Tabular Transformer with uncertainty-calibrated learning (C2T-UCL) is employed to learn clinically valid feature interactions and latent symptom dependencies among heterogeneous text features for accurate PCOS detection and to ensure reliable, low-error and confidence-aware predictions suitable for real-time decision support. The text attributes are encoded as tabular tokens and optimized using supervised contrastive and focal uncertainty losses. Three quantitative PCOS text datasets are extracted from the Kaggle and Bio-GPS repositories, which comprise 285, 250 and 541 patient records with PCOS and non-PCOS labels, and are utilized for training, testing and validation. The proposed model is assessed using Python, Scikit-learn, PyTorch, and Google Colab, and the results are compared against the baseline models such as Re-PSO, SVM-ICA, and DBMM-KPCA. **Findings:** The proposed C2T-UCL model demonstrates consistent performance with promising results across all three benchmarked datasets with 99.3% prediction accuracy, 99.2% validation accuracy, 99.6% precision, 98.8% F1-score, and a reduced RMSE of 0.6%, which indicates superior generalization and a stable PCOS diagnostic support system. **Novelty:** This research integrates a unique combination of casual-guided attention approach with supervised contrastive separation and uncertainty within a tabular transformer for efficient PCOS detection. The proposed C2T-UCL model directly learns complex feature interactions from structured text input data, which enables accurate and reliable screening process and helps practical clinical workflows for timely, cost-effective, resource-constrained and high-volume diagnostic settings, making it a promising solution for scalable PCOS diagnosis.

Keywords: PCOS Detection; Deep Learning; Tabular Transformer; Decision Support System; High-Precision System; Diagnostic Model

1 Introduction

Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder that affects women of reproductive age due to hormonal imbalance, irregular ovulation, and metabolic dysfunction, and irregular ovulation leads to long-term reproductive and metabolic dysfunction. The symptoms include menstrual irregularities, excessive androgen levels, weight gain, insulin resistance, and infertility. Diagnosing PCOS is highly dependent on clinical interpretation, as the exact cause is multifactorial, involving genetic predisposition, lifestyle factors, and disrupted endocrine signaling. Despite technological advancements and established PCOS diagnostic criteria, early and reliable detection remains challenging due to heterogeneous symptom presentation, overlapping clinical features, and class imbalance in patient screening data. Due to its rising prevalence and long-term health consequences such as type-2 diabetes, cardiovascular complications, and reproductive challenges, there is a need for intelligent automated systems that can assist the clinicians in making reliable and high-precision diagnostic decisions. Current clinical assessment methods like SVM, ensemble classifiers, CNN, vector conversion models, etc., demonstrate accuracy of detection at an early stage with fewer limitations like poor handling of class imbalance, limited interpretability, reliance on feature reduction, and high false rates under uncertain clinical conditions. In order to address these challenges, this research introduces a deep learning-based Causal-Contrastive Tabular Transformer with an uncertainty-calibrated learning model (C2T-UCL) to detect PCOS through attention mechanisms with a high learning rate to enable accurate, robust, and deployment-ready decision support systems.

A machine learning-based SMOTE stacked hybrid model for PCOS diagnosis was introduced using a dynamic feature selection method which combines the SMOTE⁽¹⁾ oversampling with a stacking ensemble method. The model evaluated six classifiers and found the best-performing classifier in handling the imbalanced PCOS dataset. In addition, the model incorporates the traditional Chinese medicine clinical indices and pulse wave parameters, which gives superior accuracy in early PCOS diagnosis compared to other methods. 87% accuracy is achieved with some limitations, including generalization and interpretability issues. A predictive PCOS detection model⁽²⁾ was developed by integrating ML with TCM to enhance PCOS diagnosis by utilizing pulse wave parameters and TCM clinical indices. Four classifiers were employed in the model, such as ET, RF, EGB and SVM; among these, SVM achieved the highest accuracy of 87.8% AUC. A total of 31 features were identified with 12 pulse parameters and 19 TCM indices. Recursive feature elimination with cross-validation techniques was employed to enhance the performance of the model. A comprehensive study⁽³⁾ was conducted on early identification of PCOS with some common associated diseases like diabetes, obesity, blood pressure and heart disease. Machine learning techniques were employed to analyze the complex interrelationships between PCOS and comorbid conditions. This multi-predictive perspective is clinically significant, as PCOS rarely occurs in isolation and often with metabolic complications. The study demonstrated that the features related to common associated conditions improve early detection of PCOS and provide a clear insight into the syndrome. The authors focused more on holistic health assessment rather than solely on research on reproductive symptoms.

A deep systematic review has been conducted by the authors examining the PCOS detection using computer-aided techniques⁽⁴⁾ by analyzing the existing literature across multiple databases, key trends, various methodological approaches and research gaps in the field. The review states that significant progress was made after employing ML techniques with several limitations, such as imbalanced datasets, feature standardization, clinical validation, interpretability and dynamic multi-modal integration. The review highlighted the importance of considering the geographic variations in PCOS occurrence while developing the advanced clinical support systems and decision-making models. A feature-driven PCOS predictor was proposed by utilizing the Reinforced Optimization model⁽⁵⁾, where the model treats the feature selection process as a decision-making problem and RLO learns the selected features through interaction with the environment. The optimization process gives an efficient mechanism to balance the exploration and exploitation. This methodology serves as an advanced optimization method which significantly boosts the PCOS prediction by identifying the most discriminative features and eliminates the redundant and irrelevant features. An exclusive Smart POSDS⁽⁶⁾ model was developed using ML techniques to create a robust detection framework by integrating various clinical and biochemical parameters to achieve reliable classification of PCOS, which demonstrates the effective handling of complex datasets by extracting feature subsets in a robust manner. The authors highlighted that the traditional ML approaches took a lead in detecting PCOS using text datasets and clinical records with fewer generalization limitations. A novel vector conversion-based PCOS detection approach was proposed using DBHMM and KPCA⁽⁷⁾. The model has done the dimensionality reduction and deep segmentation of text data. The model has the ability to automatically extract the hierarchical features from raw text data and eliminates the need for a manual feature engineering process. Here the annotated data are passed to the vector conversion process before segmentation and dimensionality reduction. This multitasking model learns the features and does parametric analysis before classifying the PCOS and normal cases. Promising results are achieved with shortcomings such as confidence-aware predictions and text encodings.

A robust clinical support system⁽⁸⁾ was introduced by employing the red deer algorithm for the feature selection process and the random forest classifier for PCOS detection. This hybrid approach demonstrates that the nature-inspired metaheuristic

algorithm effectively extracts the features from text datasets and improves classification accuracy, which helps to reduce the computational complexity. The RDA model proves that the model has a robust ability in navigating the high-dimensional feature space in clinical datasets. A web-based machine learning early PCOS screening system⁽⁹⁾ was proposed where an individual can make use of it without immediate access to specialized healthcare facilities. This method was easy to deploy and highly scalable, making the application available to a broader population for accessibility. This patient-centered healthcare approach enables self-assessment, and early intervention helps underserved populations for smooth screening and diagnosis. A deep learning CNN model⁽¹⁰⁾ was employed to detect lung cancer using DICOM datasets where deep segmentation is focused more for classification. The study provides a valuable methodological insight which is applicable for PCOS detection, particularly when using image analysis and datasets. The handling of DICOM data by 3D CNN can be adopted for an ultrasound-based PCOS screening model where ovarian morphology is crucial for this process. Ensemble methods and optimization strategies⁽¹¹⁾ are utilized to maximize the PCOS detection accuracy and minimizing the false negatives. Machine learning techniques are employed to handle the datasets and highlight the ensemble approaches, which combine the multiple base learners and help to achieve superior performance in PCOS early detection compared to individual classifiers by using the strengths of different ML algorithms. A distinctive and explainable machine learning PCOS detection framework⁽¹²⁾ was designed by emphasizing the importance of model interoperability in clinical applications. Random forest classifiers in combination with explainable methods are utilized to provide a transparent decision-making process. The XAI component allows the healthcare professional to understand the prediction system and incorporates the novel methods which help real-world clinical deployment. An effective CNN XG-Boost model⁽¹³⁾ was presented based on multimodal clinical data by integrating the deep learning-based feature extraction and classification with XG-Boost to improve the accuracy and robustness of the system. The authors have done clear preprocessing methods before feature selection and classification, which highlights the potential of hybrid models in medical diagnosis and supports AO-driven techniques for early and reliable PCOS screening and diagnosis. In addition to the above literature study, some of the specific state-of-the-art machine learning and deep learning models are analyzed for the study which is given in Table 1.

1.1 Research Gap

The background study clearly shows that various machine learning and deep learning models demonstrate the PCOS screening and detection with a multi-faceted analytical process. Some of the major features consistently emerge as important predictors for PCOS detection, which include hormonal markers (LH, FSH, testosterone, AMH), metabolic parameters (BMI, insulin resistance, glucose levels), menstrual cycle irregularities, follicle counts, lipid profiles, etc. Advanced optimization strategies are integrated for the feature selection and hyperparameter tuning process.

Despite significant progress, some of the limitations and challenges persist, such as class imbalance between PCOS and healthy records, limited availability of large, diverse and annotated datasets, inconsistent feature measurements across various clinical settings, privacy concerns of data sharing, high risk of errors, low diagnostic depth, scalability, etc. Figure 1 shows the research gap analysis and limitations of the existing models based on literature study. To overcome these limitations, a novel framework is proposed in this research to boost the performance of early PCOS detection with the help of text data.

2 Methodology

This section portrays the overall methodology and implementation of the proposed Causal-Contrastive Tabular Transformer with Uncertainty-Calibrated Learning (C2T-UCL) PCOS early detection model. The primary focus of the model is to learn complex non-linear clinical feature interactions with high robustness under class imbalance and diagnostic uncertainty. Unlike baseline models that rely on thin classifiers and feature compression, the proposed model directly learns discriminative patient representations from tabular data, which helps make the PCOS screening and detection process more reliable. In order to enhance the data quality, structured preprocessing techniques were employed, such as missing value elimination, normalization and Borderline-SMOTE (during training) to address the class imbalance. Casual guided attention and supervised contrastive learning methods are incorporated to strengthen the feature analysis and class discrimination. The UCL loss function is used for stable prediction and robust classification with minimal false rates.

The raw datasets are split into three for training, testing and validation to ensure fair assessment. The proposed C2T-UCL model is evaluated against three baseline PCOS detection and classification models includes Reinforcement-PSO⁽⁵⁾, SVM-ICA⁽⁶⁾ and DBMM-KPCA⁽⁷⁾ which provides a comprehensive benchmark for performance evaluation. Figure 2 shows the clear methodology in a step-by-step process for PCOC detection using C2T-UCL.

Table 1. Study on Specific Machine and Deep learning Models benchmarked for PCOS detection

Model	Techniques	Merits	Limitations
ML-Based PCOS Detection Model ⁽¹⁴⁾	Traditional ML classifiers such as SVM, RF, and KNN trained on clinical attributes	Simple implementation, interpretable results, moderate accuracy	Limited feature interaction learning, poor scalability, weak handling of imbalance
CNN-Based Multi-Variable PCOS Model ⁽¹⁵⁾	CNN with multi-variable feature selection and binary classification	Improved accuracy using deep feature learning, effective feature prioritization	High dependency on image quality, limited uncertainty handling
Hybrid CNN-Bi-LSTM AI Framework ⁽¹⁶⁾	CNN for spatial learning and Bi-LSTM for sequential dependency modeling	Strong temporal and contextual feature modeling	High computational cost and interoperability issues
DMA-WOA Feature Descriptor Model ⁽¹⁷⁾	Whale Optimization Algorithm with dynamic memory for feature optimization	Effective feature selection, reduced redundancy	Optimization overhead, lacks deep representation learning
Blockchain + Explainable AI PCOS System ⁽¹⁸⁾	Blockchain for data security and XAI-enabled ML classifiers	Secure data handling, improved trust and transparency	Increased system complexity, latency issues
Optimized ML with Explainable AI ⁽¹⁹⁾	Feature selection + ML classifiers with XAI interpretation	Clinically interpretable predictions, optimized feature usage	Performance depends on handcrafted features
Ultrasound Image-Based PCOS ML Model ⁽²⁰⁾	Extended ML framework using ultrasound image preprocessing	Non-invasive diagnosis, image-driven accuracy	Sensitive to noise, limited generalization
Deep Learning Tongue Biometric Model ⁽²¹⁾	CNN-based biometric feature extraction for healthcare authentication	High precision, strong pattern learning	Bio-metric oriented, relevance to PCOS web system
Deep CNN-YOLO Object Detection Framework ⁽²²⁾	CNN with YOLOv8 for object detection and analysis	High detection speed, real-time capability	Designed for vision tasks, not clinical tabular data
DL-Based PCOS Classification-Model ⁽²³⁾	Deep neural networks trained on PCOS datasets	Better performance than classical ML	Limited interpretability, imbalance sensitivity
Scleral Image-Based DL PCOS Model ⁽²⁴⁾	Deep CNN applied to scleral image analysis	Novel biometric modality, high AUC	Requires specialized imaging, low accessibility
F-Net (Follicles Net) ⁽²⁵⁾	Deep CNN for follicle segmentation and classification	Accurate follicle detection, strong spatial learning	Image-dependent, high computational demand
Deep Learning-Based PCOS Detection ⁽²⁶⁾	DNN for PCOS classification	Improved detection accuracy over traditional machine learning models; automated feature learning reduces manual intervention	Limited consideration of class imbalance; lacks uncertainty-aware prediction.
EHR-Based Machine Learning PCOS Prediction ⁽²⁷⁾	Machine learning algorithms applied to Electronic Health Records (EHR) data for predictive modeling	Utilizes real-world longitudinal health records; scalable for large healthcare systems	Focuses mainly on predictive performance without structured feature interaction modeling

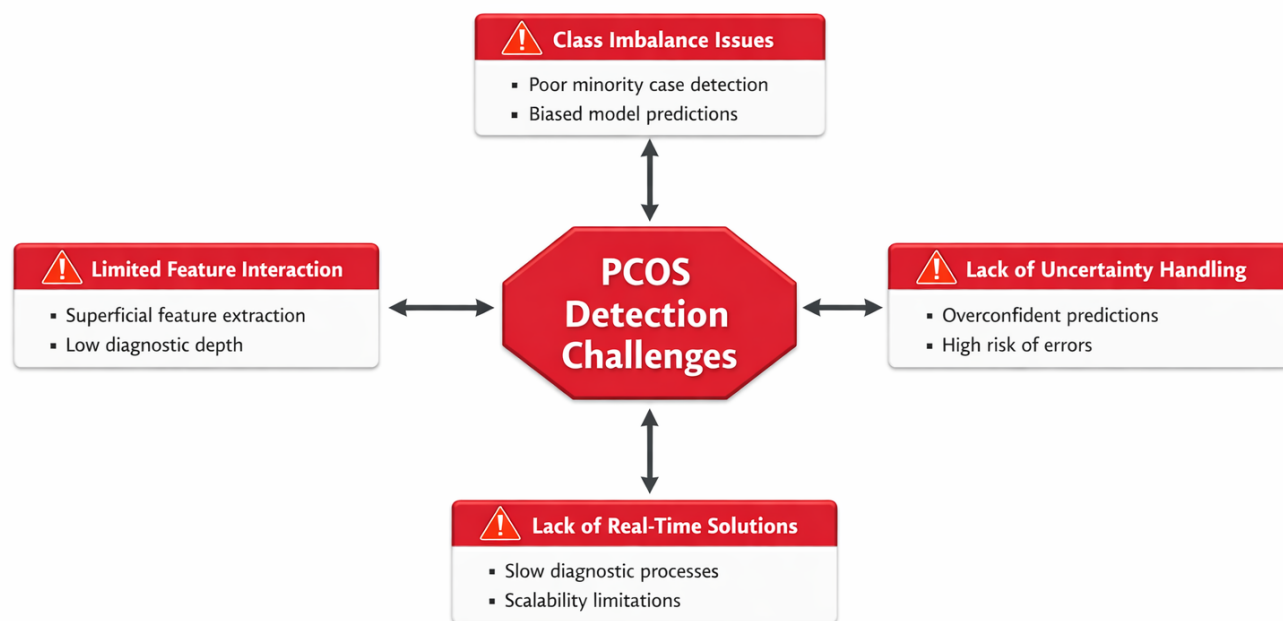


Fig 1. Research Gap Analysis

2.1 Dataset Description and Pre-Processing

The proposed model employs three benchmarked quantitative PCOS datasets, which have 285, 250, and 541 patient records, with PCOS and normal cases represented through a structural attribute under different categories. Table 2 shows clear dataset details and feature categories with data sources. These datasets reflect real-world clinical conditions, where the symptoms overlap and class imbalance between PCOS and normal patients is naturally present. Such variability makes the data more reliable for PCOS detection models. The dataset supports the proposed C2T-UCL in learning non-linear symptom dependencies, evaluation under class imbalance conditions, and validation of scalability and generalization capability. As the model requires heterogeneous feature distributions, these datasets support the model to learn the clinical interactions through transformer-based attention. Variations in sample size and feature composition enable the model to discriminate the class even in uncertainty.

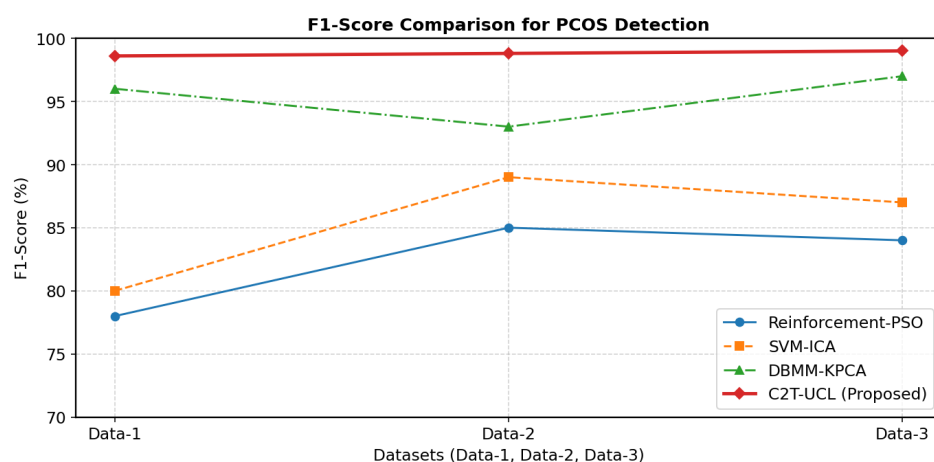


Fig 2. Research Methodology

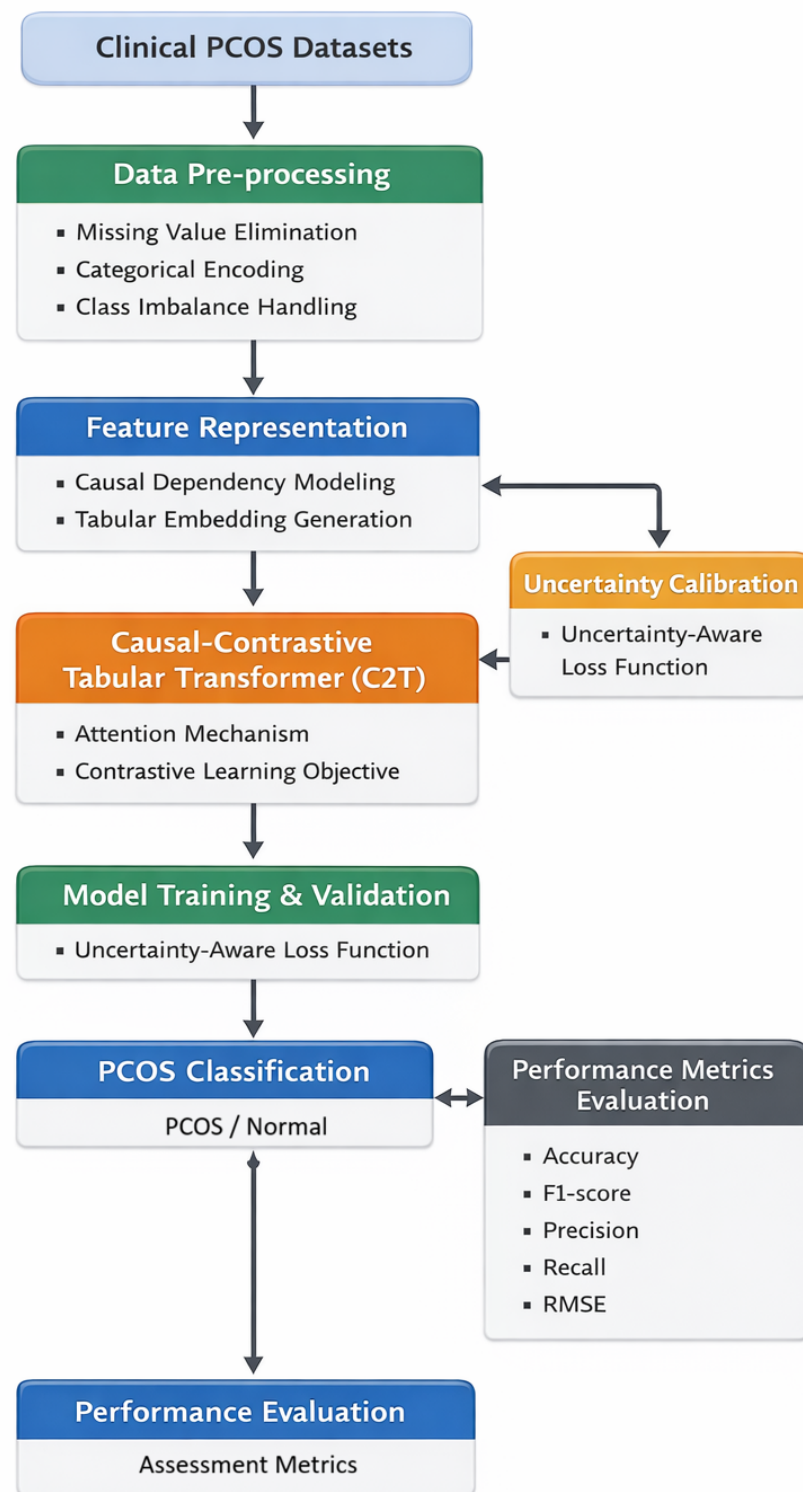


Fig 3. Pre-Processing Methods

Table 2. Dataset Description and Features

Dataset Source	Total Records	Feature Categories	Purpose
Kaggle PCOS Clinical Repository	285	Demographic, hormonal, metabolic, reproductive indicators	Serves as the primary benchmark dataset to evaluate baseline discrimination between PCOS and normal patients using structured clinical attributes.
Bio-GPS Clinical Health Records	250	Hormonal profiles, menstrual history,BMI-related features	Used to validate model robustness under moderate sample size variation and heterogeneous clinical feature distributions.
Kaggle Extended PCOS Dataset	541	Clinical symptoms, endocrine markers, lifestyle and metabolic factors	Enables large-scale generalization analysis and supports stable learning of complex feature interactions for reliable PCOS screening.
Kaggle PCOS Clinical Repository: 129 PCOS, 156 Normal, 22 Features https://www.kaggle.com/datasets/ayamoheddine/pcos-dataset Bio-GPS Clinical Health Records: 118 PCOS, 132 Normal, 20 Features http://biogps.org/dataset/tag/pcos/ Kaggle Extended PCOS Dataset: 268 PCOS, 273 Normal, 25 Features https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos			

A structured preprocessing pipeline is applied prior to model training to enhance data consistency, representation quality, and learning robustness of the proposed model, which is illustrated in Figure 3.

2.1.1. Missing Value Elimination

As the datasets are quantitative and clinical records, missing entries frequently occur due to inconsistent patient follow-ups, recording errors, and incomplete laboratory tests and readings. Such incomplete records are noisy and biased toward the transformer-based models, which depend more on contextual feature relationships. In order to clean the noisy data, the MVE method is employed as an initial data cleaning process to ensure the input data reliability. The clinical datasets are represented as feature matrix,

$$X = \{x_{ij}\}, i = 1, 2, \dots, N, j = 1, 2, \dots, F \quad (1)$$

where, N represents number of patient records, F denotes number of clinical features. Here the binary validity function v_{ij} is defined as,

$$v_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ is observed} \\ 0, & \text{if } x_{ij} \text{ is missing} \end{cases} \quad (2)$$

A record i is retained only if the proportion value of predefined features exceeds the pre-defined threshold limit T which is mathematically expressed as, $\frac{1}{F} \sum_{j=1}^F v_{ij} \geq T$. This elimination strategy ensures that, only relevant and meaningful data are taken forwarded for training process.

2.1.2. Categorical Encoding Process

Some of the input attributes cannot be directly processed by a numerical encoding model, such as menstrual regularity, symptom presence, lifestyle indicators, etc. To integrate these data into the tabular architecture, CEP is applied to convert the symbolic attributes into clear numerical representations for effective learning. The categorical feature attribute is c_j takes the input values from $C_j = \{c_1, c_2, \dots, c_K\}$. During the encoding process, each category is transformed into a numerical vector using a CE function $\phi(c_j) : C_j \rightarrow R^d$, where d denotes the attribute embedding dimension. The encoded feature vector for each patient is mathematically expressed as,

$$e_i = [\phi(c_{i1}), \phi(c_{i2}), \dots, \phi(c_{im})] \quad (3)$$

where, e_i is the categorical encoded feature vector of every patient. This complete transformation enables the attributes of different categories as tabular tokens, which are used for attention-based feature interaction learning. With the help of CEP, the C2T-UCL model effectively captures the latent representations and contextual dependencies that are critical for the PCOS detection and diagnosis process.

2.1.3. Class Imbalance Handling

The PCOS datasets commonly exhibit class imbalance, where normal/PCOS cases are outnumbered; each depends upon the population demographics. Such class imbalance is biased towards learning towards the majority class and leads to poor sensitivity for PCOS detection. In order to mitigate the class imbalance, a sample distribution strategy is used in this work. Let N_p and N_n denotes number of PCOS and normal cases respectively. Here, the imbalance ratio is expressed as, $\lambda = \frac{\max(N_p, N_n)}{\min(N_p, N_n)}$. To balance the dataset, synthetic samples are generated without affecting the original dataset and attributes and without duplicating the existing records. The augmented dataset size becomes,

$$N' = N + \alpha \cdot |N_p - N_n| \quad (4)$$

where, α denotes control equation of oversampling strength. This class handling strategy enables the supervised contrastive component to tight the intra-class cohesion and inter-class separation. In addition to the above, class-aware weighting is applied to prevent the dominance of majority samples. The class weight w_k is computed as $w_k = \frac{N}{2 \cdot N_k}$, where N_k represents total number of samples belongs to class k . Both minority PCOS case learning and uncertainty calibration during inference are improved by this class-aware weighting scheme.

2.2 Transformer-Based Tabular Tokenization and Feature Embedding

The transformer architectures are highly suitable to process the structured clinical datasets, as they have the ability to model long-range dependencies and contextual interactions. The proposed C2T-UCL model employs tabular tokenization and feature embedding to transform the heterogeneous clinical attributes into a representation that enables effective attention-based learning for PCOS diagnosis. This stage connects the pre-processed annotated data and C2T process to ensure all the clinical features contribute meaningfully to the diagnostic decision. Clinical PCOS datasets have mixed feature types, including continuous variables and encoded categorical attributes such as hormone levels, metabolic indicators, menstrual regularity, and symptom presence. Unlike traditional learning models, a transformer treats features as tokens rather than flattening the dataset into a single vector. Figure 4 shows the clear tokenization process with input, process and output. The pre-processed clinical record for a patient is expressed as,

$$x_i = [x_{i1}, x_{i2}, \dots, x_{iF}] \quad (5)$$

where, F denotes total number of clinical features. Each feature x_{ij} is mapped to token using learnable transformation function. This process converts scalar features into dense vectors which enables to process the clinical attributes in a robust manner. The feature embedding operation is represented as,

$$z_{ij} = W_e \cdot x_{ij} + b_e \quad (6)$$

where, $W_e \in R^{d \times 1}$ and $b_e \in R^d$ are learnable parameters and d is the embedding dimension. This robust transformation ensures both numerical and categorical attributes are projected into a unified latent space for further process. To address the inherent ordering, feature identity encoding is applied to retain the semantic identity of each attribute which allows the model to distinguish between the similar numeric values to different psychological measurements. The final token representation is measured as,

$$t_{ij} = z_{ij} + e_j \quad (7)$$

where, e_j denotes learnable feature embedding corresponding to j^{th} attribute. This ensures that the transformer not only recognizes the feature value, but also clinical context which is more essential for PCOS detection and screening process. The complete tokenized process is expressed mathematically as,

$$T_i = [t_{i1}, t_{i2}, \dots, t_{iF}] \quad (8)$$

where, token sequence is served as input to the transformer encoder, where casual guided attention mechanism learns the relationships of the indicators correlates with PCOS risk. As the model treats the feature attributes as tokens, premature aggregation is avoided and enables dynamic interaction learning across all dimensions.

The transformed-based tokenization has high compatibility with contrastive and uncertainty-aware learning which is employed in C2T-UCL as each feature token retains contextual identity which can effectively enforce intra-class cohesion and inter-class separation at the representation level. The tokenization feature matrix is normalized to improve the convergence and the same is expressed as, $t_{ij} = \frac{t_{ij} - \mu_j}{\sigma_j}$, where μ_j and σ_j denotes the mean and SDE of the j^{th} feature across the training data to ensure consistent scaling across heterogeneous attributes and prevent dominant features influencing the attention scores.

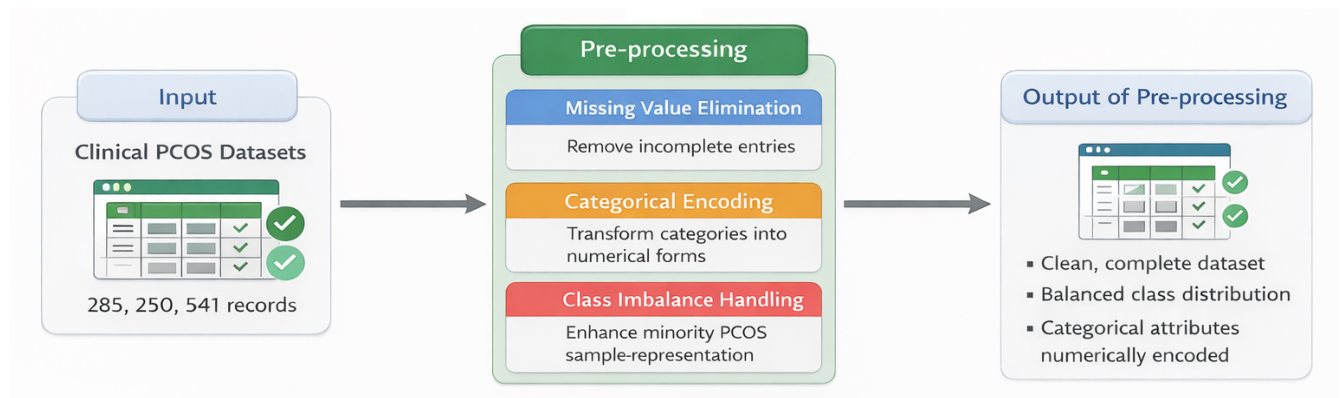


Fig 4. Tabular Tokenization for Clinical Data

2.3 Casual Guided Attention Mechanism

After structured pre-processing and tokenization stage, the clinical PCOS data is transformed into feature tokens. Though, this representation enables interaction learning; conventional self-attention mechanisms may still focus dominant features rather than clinically meaningful dependencies. In order to address this limitation, the proposed model introduced CGAM (Casual Guided Attention Mechanism) that prioritizes relationships among clinical attributes during PCOS diagnosis. Rather than arbitrary correlations, clinical indicators exhibit directional dependencies. The CGAM process resolves this by incorporating casual relevance weighting into attention computation, ensures that clinical features exert string contextual impact for further processing. Let the tokenized feature represented as $T_i = [t_{i1}, t_{i2}, \dots, t_{iF}]$, where each token t_{ij} corresponds to clinical feature obtained from transformer-tabular method. The Query, Key and Value projections are computed as,

$$Q = T_i W_Q, K = T_i W_K, V = T_i W_V \quad (9)$$

Here, the attention score is modulated using CG matrix $C \in R^{F \times F}$, where each element c_{jk} represents casual influence strength of feature j on feature k . The CGAM is defined as,

$$A_{jk} = \text{Softmax} \left(\frac{Q_j K_k}{\sqrt{d}} \cdot c_{jk} \right) \quad (10)$$

where, matrix C is learnable and refined during the CGAM training process allows the model to capture the latent features inherent in PCOS datasets without requiring the explicit casual attentions. The final feature representation is computed as,

$$h_j = \sum_{k=1}^F A_{jk} V_k \quad (11)$$

where, h_j is the final contextualized feature and each feature token is enriched with information from relevant attributes resulting in effective learning of input data by C2T-UCL model with high attention score and sensitivity to screen and detect PCOS in a robust manner.

2.4 Supervised Contrastive Learning Module

The C2T-UCL model employs supervised contrastive learning module to address the overlapping symptom patterns that occurs between affected and non-affected patients which reduces the class separability when using conventional classification loss function. SCL module enhance the discrimination between PCOS and normal cases with high TPR value. The SCL uses label information to pull representations simultaneously of the same class and push of different classes apart, where the borderline cases share partial symptom similarity with healthy patients. Let the contextual feature representation obtained from CGAM for patient i be $h_i \in R^d$, where d is the embedding dimension. The normalized similarity computation is expressed as, $h_i = \frac{h_i}{\|h_i\|}$. For each representation h_i , a set of positive samples $P(i)$ are recorded by sharing the same class label (PCOS/Normal) and negative samples $N(i)$ are eliminated to discriminate the class similarity between two feature representations. The SCL for the

given data is represented as,

$$L_{SCL}^i = \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp \frac{h_i h_p}{T}}{\sum_{a \in P(i) \cup N(i)} \exp \frac{h_i h_p}{T}} \quad (12)$$

where, T is temperature parameter which controls the smoothness of similarity class distribution. Lower T sharpens class boundaries and higher T denotes smoother separation which ensures that clinically similar features taken into consideration for further detection process. The class-aware weighting is applied to handle class imbalance which is shown above, where the weighted contrastive loss is expressed as,

$$L_{WSCL} = \frac{1}{B} \sum_{i=1}^B w_{y_i} \cdot L_{SCL}^i \quad (13)$$

This weighting mechanism ensures that, all the minority PCOS cases have a fair share of representation learning, which improves sensitivity and lowers the chance of misclassification.

2.5 Supervised Contrastive Learning Module

As the CGAM and SCL significantly improve the quality of representations and class separability, the prediction confidence is not explicitly addressed, which is one of the critical requirements in the PCOS detection process. In the uploaded data, the records often exhibit partial, conflicting, or borderline symptoms, which lead to overconfident and incorrect predictions. This limitation is addressed using UCL loss formulation by learning both predictions and their associated uncertainty, leading to safe and reliable diagnostic outputs. The latent representations produced from the tabular-transformer h_i for patient i is mapped to predictive distribution rather than single deterministic output. The UCL model estimates both predictive mean μ_i and uncertainty parameter σ_i represents the confidence associated with predictions. This process allows the model to eliminate the ambiguous clinical patterns in a robust manner. The UCL predictive likelihood is defined as,

$$p(y_i | h_i) = N(y_i | \mu_i, \sigma_i^2) \quad (14)$$

where, y_i is ground truth label and μ_i is predicted probability of PCOS and σ_i^2 captures predictive uncertainty. Based on the probabilistic formulation, the UCL for each sample is expressed as,

$$L_{UCL}^{(i)} = \frac{(y_i - \mu_i)^2}{2\sigma_i^2} + \frac{1}{2} \log(\sigma_i^2) \quad (15)$$

This loss function encourages the model to increase the uncertainty when the clinical records are insufficient which minimizes the false rate of the prediction. The final loss is calculated as,

$$L_{total} = \lambda_1 L_{WSCL} + \lambda_2 L_{UCL} \quad (16)$$

where, λ_1, λ_2 controls the uncertainty calibration and class separation. This improves accuracy and trustworthiness in PCOS prediction rates in real-time deployment by producing confidence-aware outputs with definitive predictions.

2.6 Real-Time Implementation and C2T-UCL Architecture

The proposed C2T-UCL acts as an intelligent clinical decision-making model to assist healthcare professionals in screening and detecting PCOS using an effective diagnostic method. The model operates in low-latency inference mode, which enables reliable and high-precision PCOS prediction from patient clinical records. In a real-time scenario the following steps are followed for an efficient detection process. Figure 5 shows the full C2T-UCL architecture for PCOS detection system.

- Patient data acquired from EHR, labs, or digital screening platforms
- Each data point is initially routed with a lightweight preprocessing layer to perform consistency checks, categorical normalization, and feature alignment.
- The patient data is then passed to the Transformer-Based Tabular Tokenization and Feature Embedding module to convert each feature as a token.

- Tokens are processed in parallel to enable efficient computation even if multiple patient records arrive concurrently.
- The embedded tokens are processed through CGAM, which dynamically prioritizes the meaningful feature interactions, which is essential for PCOS detection.
- SCL is activated to maintain the strong discrimination between PCOS and normal cases.
- UCL pipeline layer predicts PCOS with an associated confidence score.

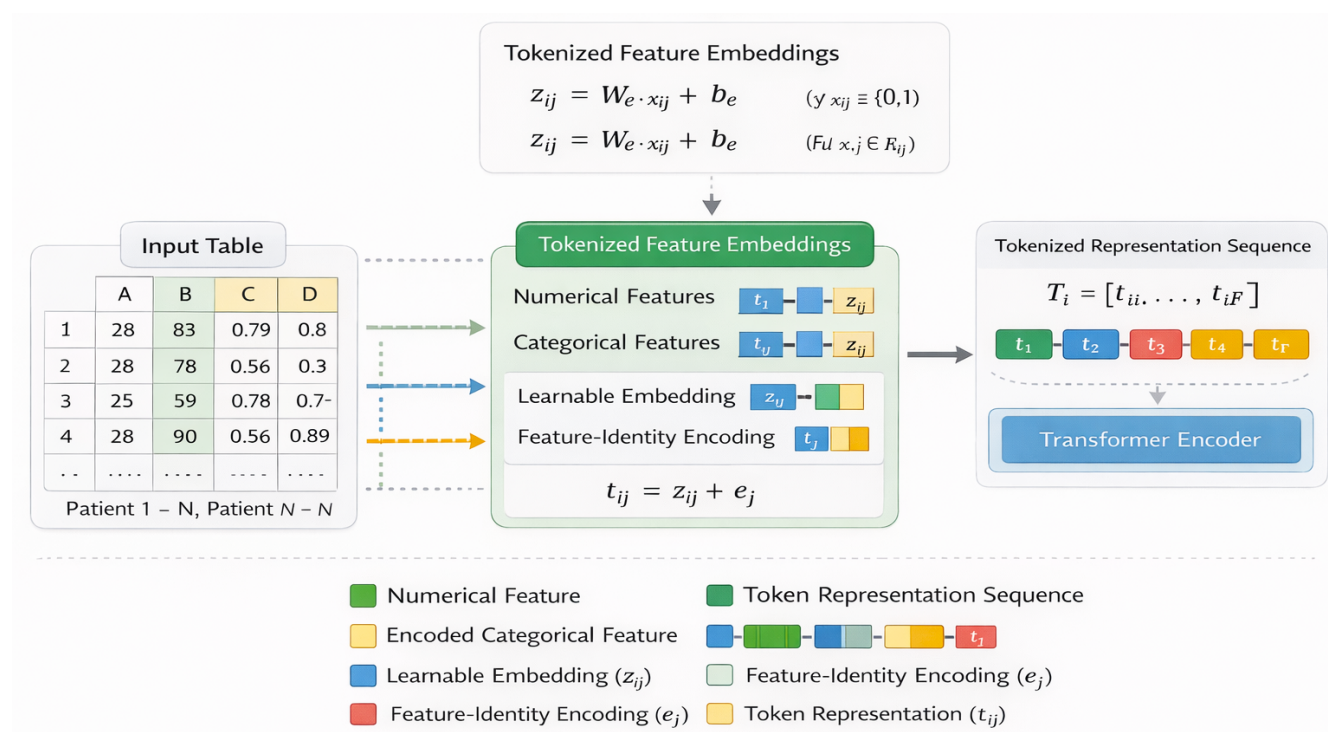


Fig 5. C2T-UCL Architecture and Flow Diagram

Real-time scenario: Consider a patient undergoing a routine screening process; the clinician enters the attributes, such as age, BMI, menstrual regularity, androgen levels, insulin markers, and reported symptoms, into the hospital system. Upon submission, the C2T-UCL system processes the data in real time, where the transformer assigns the contextual importance to features like androgen levels and irregular menstruation, while contrastive representations position the patient closer to PCOS. The UCL model detects uncertainty and outputs the PCOS-positive prediction with an associated confidence score. The system flags this case as PCOS-likely with moderate uncertainty, which prompts the clinician to recommend the patient for additional tests rather than immediate intervention. This makes the system more proactive, reliable, and scalable; the diagnostic support process effectively suits high-volume, resource-constrained, and time-sensitive healthcare environments.

2.7 Model Training, Validation and Inference Strategy

During training, the mini-batch gradient descent method is used for parameter optimization to make the system efficiently learn and extract data from a large-volume dataset. Each feature is converted as tokens and processed concurrently. The overall training loss is calculated using the UCL loss function and prevents overfitting. The training integrated three core objectives, such as casual-guided representation learning through attention layers, SCL to enhance class separability, and UCL to regulate prediction confidence. Validation is performed at the end of each epoch using unseen validation data with performance metrics. The best validation performance is retained for the inference point. The model functions in a feed-forward, low-latency mode during inference, which is suitable for the real-time screening procedure to guarantee precise classification with a high confidence score. The cases with low-confidence score will be marked for additional examination and analysis. Table 3, 4 and Figure 6 illustrate the dataset split ration, training and validation loss.

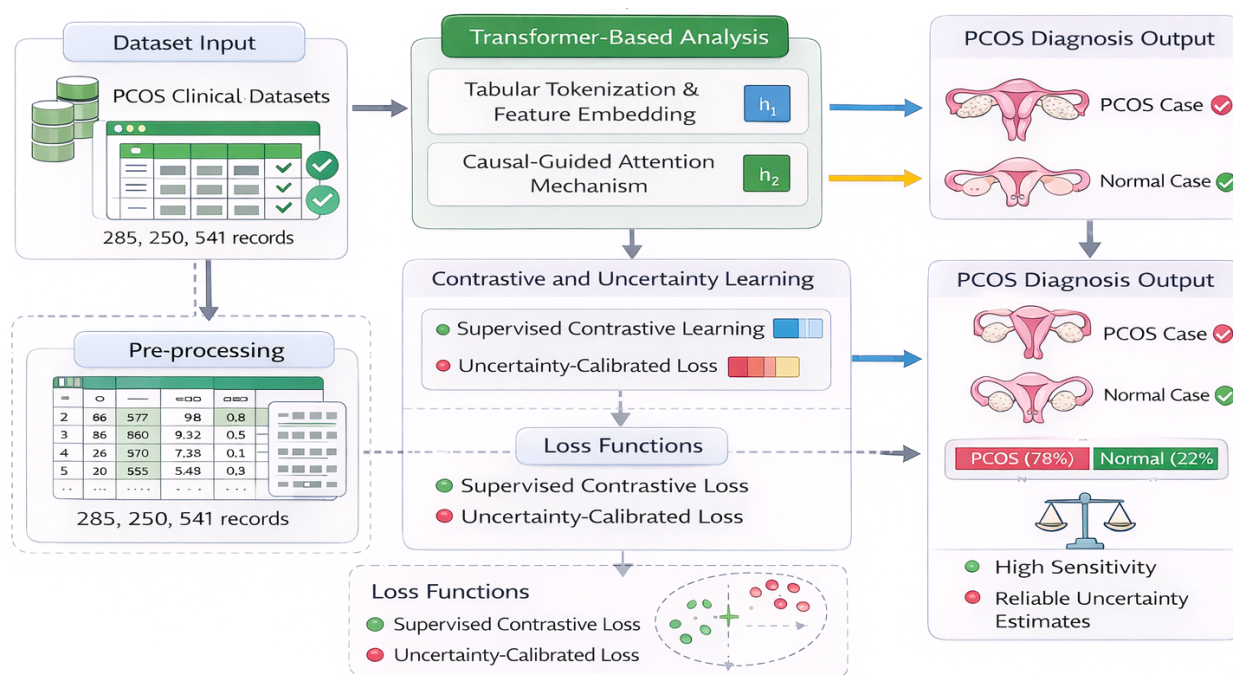
Table 3. Dataset Split Ratio and Purpose

Dataset	Total Records	Training Set (70%)	Validation Set (15%)	Testing Set (15%)	Purpose
Dataset-1	285	200	43	42	Core training and baseline performance evaluation
Dataset-2	250	175	38	37	Generalization validation across feature variation
Dataset-3	541	379	81	81	Large-scale robustness and stability assessment

Training Set - For Parameter Optimization
Validation Set - Hyperparameter tuning and Early Stopping
Testing Set - Reserved for Final Performance Testing

Table 4. Training and Validation Loss (Sample 10 Epochs)

Metric/ Epoch	1	2	3	4	5	6	7	8	9	10
Training Loss	0.420	0.310	0.240	0.180	0.130	0.090	0.060	0.045	0.032	0.025
Validation Loss	0.450	0.340	0.270	0.210	0.150	0.110	0.080	0.055	0.041	0.033

**Fig 6.** Training and Validation Loss on Epochs (100 Epochs)

2.8 Confusion Matrix and Performance Values

Figure 7 shows a clear, comprehensive evaluation for PCOS classification of the proposed C2T-UCL model across three benchmarked datasets. The model's high discriminative capability is visible in the matrix results.

Dataset 1: Out of 129 cases, 127 are identified correctly with 2 FNR, while normal cases show only a single false positive, which portrays a high sensitivity rate.

Dataset 2: Zero FPR and only 1 PCOS case misclassified as normal highlights more precision and reliability.

Dataset 3: 1 FPR and 1 FNR confirm strong generalization and stability under large-scale data.

Overall, the model witnessed very low false rates, which reduces the diagnosis risks, and the false rates enable high reliability and precision outcomes.

C2T-UCL Step by Step Algorithm

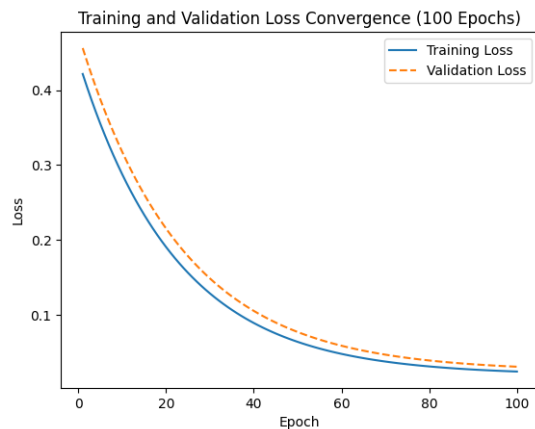


Fig 7. Actual and Predicted Values on the given datasets (Dark Red → High Correct Predictions (TP, TN) / Light yellow → Very Low Misclassifications (FP, FN))

Table 5. Real-Time Parameter Settings for Clinical Implementation

Parameters	Value / Settings	Description
A. Data & Pre-processing		
Datasets	3 tabular datasets (N=285, 250, 541)	Quantitative clinical records with PCOS/Normal labels; used for training/validation/testing and cross-dataset robustness checks.
Split strategy	Stratified 70% / 15% / 15%	Preserves class proportions in train/validation/test; avoids leakage by fitting preprocessing only on training data.
Missing value rule (τ_{miss})	Drop records with < 90% observed features	Removes incomplete clinical entries that can destabilize attention patterns and uncertainty estimation.
Numerical scaling	Z-score normalization	Standardizes continuous attributes to comparable scale; improves optimizer stability and attention score conditioning.
Categorical encoding	One-hot (low-card) / ID+Embedding (high-card)	Transforms non-numeric attributes into numeric form; supports transformer tokenization while preserving feature identity.
Imbalance handling	Borderline-SMOTE (train only)	Augments minority-class samples near decision boundary to improve PCOS sensitivity while reducing bias toward the majority class.
Borderline-SMOTE k	$k_{neighbors} = 5$	Controls neighborhood used to synthesize minority samples; tuned by validation to limit noise injection.
B. Transformer-Based Tokenization & Embedding		
Embedding dimension (d)	64	Size of each feature token vector; balances expressiveness and compute cost for clinical tabular data.
Feature-identity embeddings	Learned $e_j \in \mathbb{R}^d$	Adds feature semantics (which feature), preventing confusion between similar numeric ranges from different measurements.
Token sequence length	F tokens per patient	One token per clinical attribute; enables explicit feature–feature interaction modeling via self-attention.

Continued on next page

Embedding dropout	0.10	Regularizes token representations; reduces overfitting on small-to-medium clinical datasets.
C. C2T Transformer Backbone		
Encoder layers (L)	4	Depth of transformer encoder; captures higher-order feature interactions without excessive latency.
Attention heads (H)	8	Multi-head attention learns diverse interaction subspaces (hormonal, metabolic, symptom clusters).
FFN hidden size	256	Nonlinear mixing capacity in transformer feed-forward blocks.
Activation	GELU	Smooth nonlinearity commonly used in transformer encoders; supports stable gradients.
Dropout (encoder)	0.10	Prevents co-adaptation in attention/FFN layers; improves generalization.
D. Causal-Guided Attention		
Causal guidance matrix (C)	$Learnable\ C \in \mathbb{R}^{F \times F}$	Modulates attention weights to emphasize clinically plausible dependencies and suppress spurious correlations.
Guidance strength (α)	1.0 (tunable 0.5–2.0)	Scales the influence of causal guidance on attention logits; tuned on validation to avoid over-constraining learning.
E. Supervised Contrastive Learning		
Projection head	MLP $\rightarrow$ 128-d (normalized)	Maps encoder outputs to a contrastive space for class-wise clustering (PCOS vs Normal).
Temperature (τ)	0.07	Controls sharpness of similarity distribution; smaller τ enforces tighter clusters and clearer margins.
Batch sampling	Class-aware batching	Ensures each batch contains positives/negatives for supervised contrastive pairs; improves stability under imbalance.
F. Uncertainty-Calibrated Learning (UCL)		
Uncertainty head	Predict $\log \sigma^2$	Outputs an uncertainty scale for each sample; discourages overconfident errors and supports confidence-aware screening.
Uncertainty regularization	Clamp $\log \sigma^2$ to $[-13.8, 0]$	Prevents numerical instability (σ^2 too small/large) and avoids trivial inflation of uncertainty.
Loss weights	$\lambda_1 = 1.0$ (SCL), $\lambda_2 = 0.5$ (UCL)	Balances class separation with calibration; tuned on validation to maximize F1 while keeping uncertainty meaningful.
G. Training & Validation		
Optimizer	Adam-W	Handles sparse gradients and regularization well for transformer training on tabular data.
Learning rate (η)	1e-4	Stable default for transformer fine-tuning; adjusted via validation if convergence is slow/unstable.
Weight decay	1e-2	Regularizes weights to reduce overfitting while preserving representation quality.
Batch size (B)	64	Balances gradient stability and GPU memory constraints; may be reduced for larger data.
Epochs (E)	Up to 50	Training budget; typically converges earlier with early stopping.

Continued on next page

Early stopping	Patience = 10 epochs	Stops when validation loss/metrics stop improving; prevents overfitting.
LR schedule	Cosine decay + 5-epoch warmup	Improves stability in early epochs and refines convergence near optimum.
Random seed	42	Ensures reproducibility of splits, SMOTE sampling, and initialization.
H. Inference & Real-Time Deployment		
Decision threshold	$p(PCOS) \geq 0.50$	Default binary decision threshold; can be shifted to reduce FN in screening contexts.
Low-confidence flag	Confidence < 0.60 OR σ above threshold	Routes ambiguous cases for clinician review; supports safe deployment.
Serving mode	Batch=1 / micro-batch	Supports real-time single-patient inference or small batches for throughput.

2.9 Performance Evaluation

In order to assess the effectiveness, robustness, and clinical reliability of the proposed C2T-UCL model, a comprehensive assessment is conducted with five performance metrics, such as accuracy, precision, recall, F1-score, and root mean square error (RMSE). Based on real-time parameter settings for clinical implementation presented in Table 5, the evaluation is performed using three heterogeneous datasets with small-scale and large-scale diagnostic scenarios. The trained model is tested using unseen data, and experiments are conducted under consistent parameter settings for comparative analysis. The following notations are used in performance metrics,

- **TP**: True Positives (PCOS correctly predicted as PCOS)
- **TN**: True Negatives (Normal correctly predicted as Normal)
- **FP**: False Positives (Normal incorrectly predicted as PCOS)
- **FN**: False Negatives (PCOS incorrectly predicted as Normal)
- y_i : True label of sample i
- $\hat{y}_i$: Predicted output (probability or class)
- **N** : Total number of test samples

$$\text{Accuracy} = \frac{(TPR + TNR)}{(TPR + TNR + FPR + FNR)} \times 100 \quad (17)$$

$$\text{Recall (Sensitivity)} = \frac{TPR}{(TPR + FNR)} \times 100 \quad (18)$$

$$\text{Precision} = \frac{TPR}{(TPR + FNR)} \quad (19)$$

$$\text{F1-Score} = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (20)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (21)$$

The performance metrics show the model's potential in screening, detecting, and classifying the PCOS in an effective manner. Here, accuracy captures overall effectiveness of PCOS detection, precision reduces false PCOS alarms, recall minimizes missed PCOS cases, the F1-score handles symptom overlap and class imbalance, and RMSE evaluates confidence and prediction stability.

3 Results and Discussions

This section presents a comprehensive analysis and experimental results of the proposed C2T-UCL model for the PCOS screening and detection process. The model is evaluated systematically on how the system translates the C2T-UCL into measurable diagnostic performance across benchmarked clinical HER datasets. Python is used for training, validation, and testing under consistent parameter settings. Comparative analysis is done with baseline PCOS detection models such as Reinforcement-PSO⁽⁵⁾, SVM-ICA⁽⁶⁾, and DBMM-KPCA⁽⁷⁾ to highlight the superiority of the proposed model in terms of accuracy, robustness, and confidence-aware PCOS diagnosis. The contribution of each component is tested under an ablation study to examine the impact of CGAM, SCL, and UCL on overall performance shown in Table 7 and 8. An impact analysis is done to analyze the model on clinical reliability and decision confidence. Computational complexity assessment is done to evaluate the training convergence and inference feasibility in real time. Table 6 and Figures 8- 12 show the performance results of the proposed model and provide meaningful insights on the scalability and clinical applicability of the diagnostic model.

Table 6. Findings and Comparative Analysis (Existing and Proposed)

Dataset	Techniques	Prediction Accuracy	Validation Accuracy	RMSE	Precision	F1-Score
Data 1	Reinforcement-PSO	85	88	45	83	78
	SVM-ICA	88	89	46	86	80
	DBMM-KPCA	97	97	2	96	96
	C2T-UCL [Proposed]	99	99	0.7	99	98
Data 2	Reinforcement-PSO	92	86	57	84	85
	SVM-ICA	94	94	58	89	89
	DBMM-KPCA	96	95	4	96	93
	C2T-UCL [Proposed]	99	99	0.6	99	98
Data 3	Reinforcement-PSO	88	91	59	87	84
	SVM-ICA	94	95	60	90	87
	DBMM-KPCA	98	98	1	99	97
	C2T-UCL [Proposed]	99	99	0.5	99	99

3.1 Prediction Accuracy

The accuracy metric is used to evaluate the model's overall performance and determine whether the proposed approach accurately classifies PCOS and normal cases across various datasets. Given that CGAM and the tokenization method are integrated to effectively capture the complex features and non-linear relationships in PCOS data, C2T-UCL consistently achieves high accuracy across epochs. The results are compared with baseline models Reinforcement-PSO⁽⁵⁾, SVM-ICA⁽⁶⁾, and DBMM-KPCA⁽⁷⁾, which rely on shallow decision methods. The adaptability of existing models is limited compared to the C2T-UCL model, as it has superior classification through various layers like SCL, UCL, etc. Dynamic feature interactions are done at multiple levels of abstractions, helping the model to outperform in all metrics. High accuracy shows the robustness of the model with generalization rather than overfitting and confirms that the new architecture scales effectively and learns latent representations from diverse data, making it reliable and high precision under real-world scenarios. 99% accuracy is recorded in C2T-UCL compared to baseline detection models.

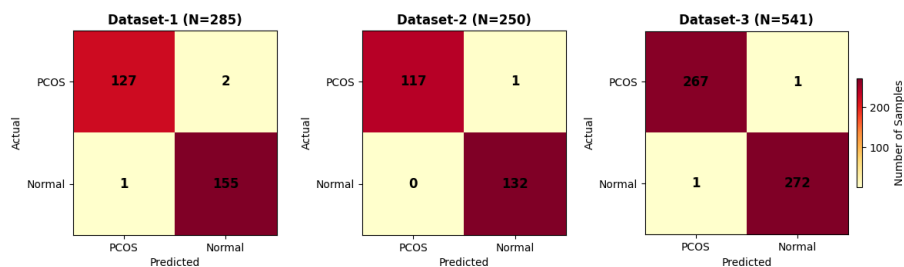


Fig 8. Prediction Accuracy Analysis

3.2 Validation Accuracy

The model's generalization capability and overfitting resistance are measured using validation accuracy. The SCL and UCL methods learn the latent representations effectively. The feature tokens of each patient are validated with data and regularized with CGAM to perform confidence-aware predictions. The training patterns are captured by the model to learn the clinical representations to support stable learning through attention-based aggregation. The proposed model maintains diagnostic reliability when new datasets are deployed with new populations, which proves the suitability for a scalable and safe PCOS screening process. 99% validation accuracy is achieved by C2T-UCL model without any overfitting and false rates. All three datasets show consistent performance in detecting PCOS with high validation accuracy.

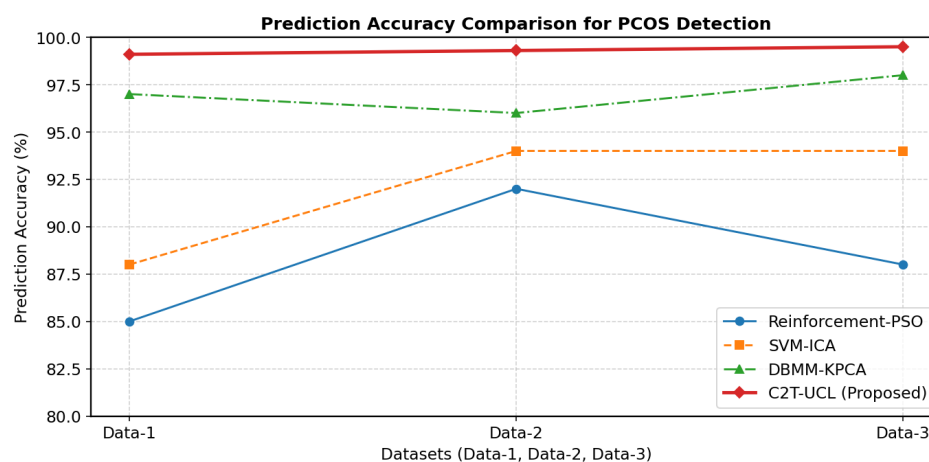


Fig 9. Validation Accuracy Analysis

3.3 RMSE Analysis

The prediction stability and confidence behavior of the proposed model are measured using RMSE. The model produces accurate predictions that are consistently close to true labels, which confirms the effectiveness of UCL in regulating the prediction confidence. Baseline models attain high RMSE due to shallow methods and unstable probability and incorrect prediction. The proposed C2T-UCL addresses the limitation by learning both prediction and uncertainty to classify classes with high sensitivity and precision. Prediction errors are minimized in C2T-UCL, which proves that real-time diagnostic reliability is proven with high true positive rates. As a result, the model shows high interpretability and clinical trustworthiness in PCOS screening, detection, and diagnosis.

3.4 Precision

The precision analysis portrays how the proposed model identifies PCOS cases among predicted positives and reduces the risk of false alarms. The C2T-UCL model rarely misclassifies the normal cases as PCOS positive, which confirms that the decision boundaries are well-calibrated. Baseline models record low precision due to class imbalance and overgeneralization. The proposed C2T-UCL overcomes the limitations by structuring the features as tokens using tokenization and latent learning through SCL and ensures tight clustering of true PCOS cases while maintaining clear separation from normal cases. Furthermore, CGAM reduces the spurious correlations and regulates all the dominant features to contribute to the prediction model. High precision is achieved across all datasets by the proposed model, making it suitable for large-scale screening with minimal follow-ups and real-time environments.

3.5 F1-Score

The harmonic balance between precision and recall is measured using the F1 score, where the symptom overlap and class imbalance are measured and controlled in the PCOS detection process. The superior F1 score maintains high sensitivity and specificity simultaneously. An optimal trade-off between affected and normal patients is clearly discriminated to avoid false diagnoses. Existing models attain low F1-scores and harmonic balance as they use collaborated feature analysis with high

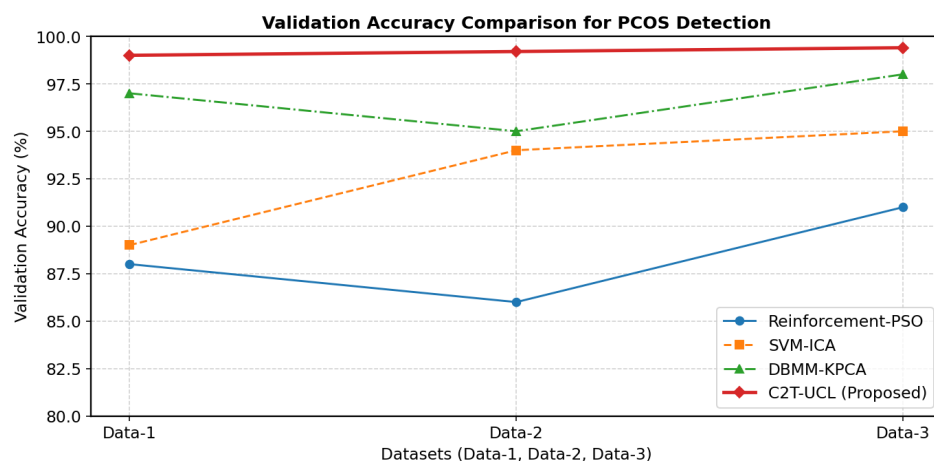


Fig 10. RMSE Analysis

dataset-dependent variability. The proposed C2T-UCL captures subtle features without relying on dependent variability, which indicates the stability of the model aligns closely with real clinical reasoning. The model achieved a 99% F1 score, which outperforms all existing models and reinforces the reliability for practical healthcare deployment. The model fits for diverse datasets, which are tested under impact and complexity analysis, and high performance is recorded.

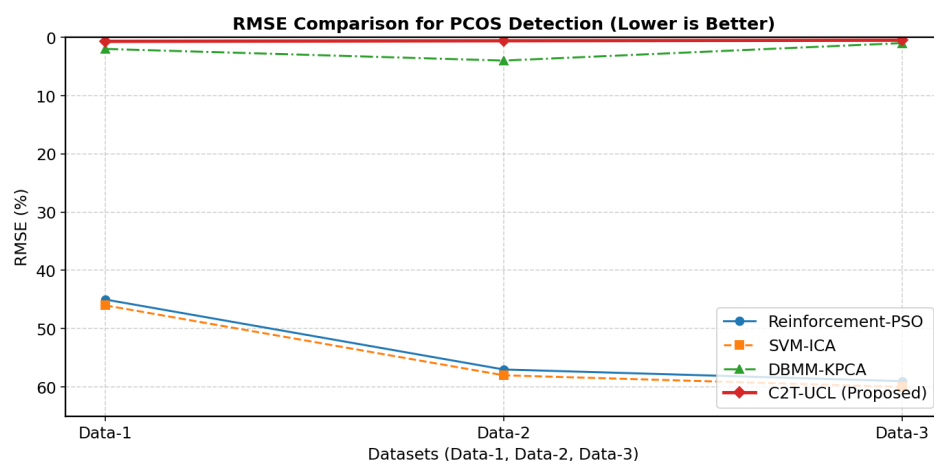


Fig 11. Precision Analysis

3.6 Overall Discussion

The experimental results demonstrate that the proposed C2T-UCL model achieves high predictive performance compared to baseline machine learning and deep learning models. Standard DNN and classical ML models lack robustness under class imbalance and heterogeneous clinical attributes^{(26), (27)}. To overcome these limitations, this transformer-based model enables structured feature interactions which aligns the recent advances in attention driven representation learning. The dynamic integration of SCL enhances the class separability which addresses the limitations of HER based predictive systems. UCL provides the superior controlled confidence estimation which was not explicitly incorporated in the baseline studies. These integrations made the model outperform in accuracy, precision and RMSE across diverse datasets leads to deployment ready PCOS detection support systems.

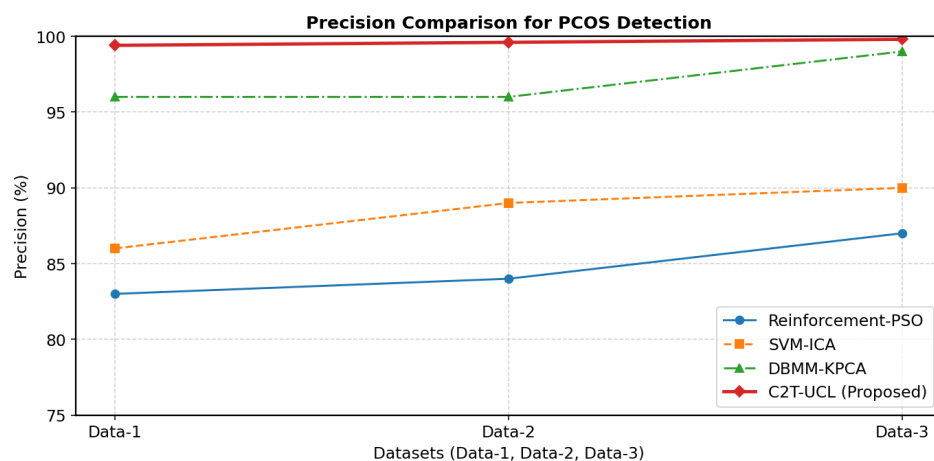


Fig 12. F1 Score Analysis

Table 7. Ablation Study and Impact Analysis

Model Variant	Prediction Accuracy (%)	Precision (%)	Key Observation (Impact)
C2T-UCL (Full Proposed)	99.3	99.6	Best overall; stable and confidence-aware PCOS decision support
Without Causal-Guided Attention	98.6	98.9	More spurious feature links; higher FN/FP in borderline cases
Without Supervised Contrastive Learning (use only classifier loss + UCL)	98.4	98.7	Weaker class separation; embedding space less structured
Without UCL (no uncertainty calibration; standard classification loss)	98.7	98.8	Overconfident errors rise; stability reduces (RMSE increases)
Without Class-Imbalance Handling (no Borderline-SMOTE / no class weighting)	97.9	98.2	Minority PCOS sensitivity drops; increased FN in skewed splits
Without Feature-Identity Embedding (token value embedding only)	98.2	98.5	Feature semantics blurred; attention less clinically consistent

Table 8. Complexity Analysis

Model Variant	Parameters (×)	Training Time per Epoch (s)	Complexity Note
Full C2T-UCL	1.00×	7.6	Full pipeline: causal attention + contrastive head + UCL head
Without Causal Guidance	0.99×	7.2	Removes causal weighting/gating; slightly cheaper attention
Without Contrastive Module	0.92×	6.8	Drops projection head + pairwise similarity computations
Without UCL	0.97×	7.1	Removes uncertainty head and its loss terms
Without Imbalance Handling	1.00×	6.9	Less data augmentation; faster epochs but poorer minority learning
Without Feature-Identity Embedding	0.98×	7.3	Slightly fewer embedding parameters

4 Conclusion

This research presents a hybrid transformer-based C2T-UCL model to predict PCOS at an early stage with high precision and reliability using three benchmarked and structured clinical PCOS data extracted from Bio-GPS and Kaggle repositories. To overcome the limitations of existing models, the new model is proposed to enhance class discrimination, feature analysis, and confidence-aware decision-making. Three heterogeneous datasets are used for comprehensive evaluation, and the experimental results demonstrate 99.3% prediction accuracy, 99.2% validation accuracy, 99.6% precision, a 98.8% F1-score, and a reduced RMSE of 0.6% with systematically validated results. Consistent performance is recorded as the model uses tabular tokenization, SCL, CGAM, and UCL methods for learning, and class separation by regulating the prediction confidence makes the model suitable for real-time clinical decision support. Strong generalization and interpretability are demonstrated across diverse datasets due to its modular architecture, and confidence-aware output showcases a potential system for PCOS screening and detection processes in resource-constrained environments. Though the model shows superior performance, the C2T-UCL model relies on structured clinical data and does not work with imaging like 2D and 3D datasets and longitudinal patient records. Future work can integrate multi-modal datasets and dynamic temporal analysis to enhance the clinical adaptability and diagnostic coverage.

5 Acknowledgements

The author express sincere gratitude to Dr SNS Rajalakshmi College of Arts & Science for providing infrastructure support and all those who contributed to the successful completion of this research work. Sathiyapriya Jagadeesan played a major role in carrying out the research, including dataset collection, literature review, system implementation, and drafting of the manuscript. Praveena Marannan, research guide, has done a critical review, document checking, correcting, finalizing to improve the quality and clarity of the manuscript, which made the research meaningful and complete.

6 List of Abbreviations

- C2T-UCL – Causal-Contrastive Tabular Transformer with Uncertainty-Calibrated Learning
- EHR – Electronic Health Record
- CDSS – Clinical Decision Support System
- TT – Tabular Tokenization
- FE – Feature Embedding
- CGAM – Causal-Guided Attention Module
- SCL – Supervised Contrastive Learning
- UCL – Uncertainty-Calibrated Learning
- BCE – Binary Cross-Entropy
- SMOTE – Synthetic Minority Over-sampling Technique
- B-SMOTE – Borderline-SMOTE
- RMSE – Root Mean Square Error
- AUC-ROC – Area Under the Receiver Operating Characteristic Curve

7 Declaration

The authors declare that this study is original, independently conducted, and hasn't been submitted for publication anywhere else. Implementation tools include MATLAB, Python, and PyTorch are used for experimental analysis. Additionally, Mermaid is used for graphical abstracts, and Grammarly and Open-AI are used for language refinement.

References

- 1) Kumari R, Singh J, Gosain A. SmS: SMOTE-stacked hybrid model for diagnosis of polycystic ovary syndrome using feature selection method. *Expert Systems with Applications*. 2023;225:120102. <https://doi.org/10.1016/j.eswa.2023.120102>.
- 2) Lim J, Li J, Feng X, Feng L, Xiao X, Xia Y, et al. Machine learning-based evaluation of application value of traditional Chinese medicine clinical index and pulse wave parameters in the diagnosis of polycystic ovary syndrome. *European Journal of Integrative Medicine*. 2023;64:102311. <https://doi.org/10.1016/j.eujim.2023.102311>.
- 3) Aggarwal S, Pandey K. Early identification of PCOS with commonly known diseases: Obesity, diabetes, high blood pressure and heart disease using machine learning techniques. *Expert Systems with Applications*. 2023;217:119532. <https://doi.org/10.1016/j.eswa.2023.119532>.

- 4) Suha SA, Islam MN. A systematic review and future research agenda on detection of polycystic ovary syndrome (PCOS) with computer-aided techniques. *Heliyon*. 2023;9(10):e20524. <https://doi.org/10.1016/j.heliyon.2023.e20524>.
- 5) Reka S, Praba TS, Manchala KK. Comprehensive Feature-Driven PCOS Predictor: A Reinforcement Learning-Based Binary Equilibrium Optimization Approach. *International Journal of Computational Intelligence Systems*. 2025;18:188. <https://doi.org/10.1007/s44196-025-00931-3>.
- 6) Tiwari S, Kane L, Koundal D, Jain A, Alhudhaif A, Polat K, et al. SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning. *Expert Systems with Applications*. 2022;203:117592. <https://doi.org/10.1016/j.eswa.2022.117592>.
- 7) Jagadeesan S, Marannan P. Vector Conversion Based PCOS Detection in Data Segmentation using Multi Task Learning by Dynamic Deep Learning Architecture. *Biotechnology Research Asia*. 2025;22(1). <https://doi.org/10.13005/bbra/3355>.
- 8) Sreejith S, Nehemiah HK, Kannan A. A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier. *Healthcare Analytics*. 2022;2:100102. <https://doi.org/10.1016/j.health.2022.100102>.
- 9) Rahman MM, Islam A, Islam F, Zaman M, Islam MR, Sakib MSA, et al. Empowering early detection: A web-based machine learning approach for PCOS prediction. *Informatics in Medicine Unlocked*. 2024;47:101500. <https://doi.org/10.1016/j.imu.2024.101500>.
- 10) Eldho KJ, Nithyanandh S. Lung Cancer Detection and Severity Analysis with a 3D Deep Learning CNN Model Using CT-DICOM Clinical Dataset. *Indian Journal of Science and Technology*. 2024;17(10):899–910. <https://doi.org/10.17485/IJST/v17i10.3085>.
- 11) Panjwani B, Yadav J, Mohan V, Agarwal N, Agarwal S. Optimized Machine Learning for the Early Detection of Polycystic Ovary Syndrome in Women. *Sensors*. 2025;25(4):1166. <https://doi.org/10.3390/s25041166>.
- 12) Khanna VV, Chadaga K, Sampathila N, Prabhu S, Bhandage V, Hegde GK. A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome. *Applied System Innovation*. 2023;6(2):32. <https://doi.org/10.3390/asi6020032>.
- 13) Pawar S, Dhane P, Shelke D, Dheple P, Doshi N. PCOS Detection using Hybrid CNN-XGBoost Model – A Multimodal Data Approach. *Biomedical and Pharmacology Journal*. 2025;18(3). <https://doi.org/10.13005/bpj/3252>.
- 14) Hdaib D, Almajali N, Alquran H, Mustafa WA, Al-Azzawi W, Alkhayyat A. Detection of Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms. *Proceedings of the 5th International Conference on Engineering Technology and its Applications (IICETA)*. 2022;p. 532–536. <https://doi.org/10.1109/IICETA54559.2022.9888677>.
- 15) Sathiyapriya J, Praveena M. Enhanced PCOS Detection Using CNN Multi Variable Feature Selection and Binary Classification for Precise Diagnosis. *Proceedings of the 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*. 2025;p. 1349–1354. <https://doi.org/10.1109/ICDICI66477.2025.11134990>.
- 16) Selvam N. AI-Driven Indian SLR Using Hybrid CNN-BiLSTM Architecture for Divyangjan. *International Journal of Emerging Science and Engineering*. 2025;14(1):14–25. <https://doi.org/10.35940/ijese.L2628.14011225>.
- 17) Amissah DK, Boante LM, Mensah S, Owusu E, Appati JK. Dynamic Memory-Augmented Whale Optimization Algorithm (DMA-WOA) as Feature Descriptor for Polycystic Ovary Syndrome Detection. *Applied AI Letters*. 2026;7(1):e70016. <https://doi.org/10.1002/ail2.70016>.
- 18) Jaganathan G, Natesan S. Blockchain and Explainable-AI Integrated System for Polycystic Ovary Syndrome (PCOS) Detection. *PeerJ Computer Science*. 2025;11:e2702. <https://doi.org/10.7717/peerj-cs.2702>.
- 19) Elmannai H, El-Rashidy N, Mashal I, Alohali MA, Farag S, El-Sappagh S, et al. Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence. *Diagnostics*. 2023;13(8):1506. <https://doi.org/10.3390/diagnostics13081506>.
- 20) Suha SA, Islam MN. An Extended Machine Learning Technique for Polycystic Ovary Syndrome Detection using Ovary Ultrasound Image. *Scientific Reports*. 2022;12:17123. <https://doi.org/10.1038/s41598-022-21724-0>.
- 21) Indhumathi G, Anil PS, Posiyya A, Suresh HR, Navaneethan S. Deep Learning-Based Tongue Biometrics for Secure Authentication in IoT-Driven Healthcare Systems. *Proceedings of the International Conference on Smart & Sustainable Technology (INCSST)*. 2025;p. 1–6. <https://doi.org/10.1109/INCSST64791.2025.11210319>.
- 22) Nithyanandh S. Object Detection & Analysis with Deep CNN and YOLOv8 in Soft Computing Frameworks. *International Journal of Soft Computing and Engineering*. 2025;14(6):19–27. <https://doi.org/10.35940/ijsc.E3653.14060125>.
- 23) Gupta K, Prasad R. Polycystic Ovary Syndrome Detection using Deep Learning. 2023;p. 1465–1468. <https://doi.org/10.1109/IC3I59117.2023.10397615>.
- 24) Lv W, Song Y, Fu R, Lin X, Su Y, Jin X, et al. Deep Learning Algorithm for Automated Detection of Polycystic Ovary Syndrome Using Scleral Images. *Frontiers in Endocrinology*. 2022;12. <https://doi.org/10.3389/fendo.2021.78987>.
- 25) S S, Umapathy S, Alhajlah O, Almutairi F, Aslam S, et al. F-Net: Follicles Net an Efficient Tool for the Diagnosis of Polycystic Ovarian Syndrome Using Deep Learning Techniques. *PLOS ONE*. 2024;19(8):e0307571. <https://doi.org/10.1371/journal.pone.0307571>.
- 26) Bairagi VK, et al. Polycystic Ovarian Syndrome (PCOS) Detection Through Deep Learning. *International Journal of Engineering Trends and Technology*. 2024;72(10):159–170. <https://doi.org/10.14445/22315381/IJETT-V72I10P116>.
- 27) Zad Z, Jiang VS, Wolf AT, Wang T, Cheng JJ, Paschalidis IC, et al. Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. *Frontiers in Endocrinology*. 2024;15:1–14. <https://doi.org/10.3389/fendo.2024.1298628>.