A Comprehensive Review of Machine Learning and Deep Learning Approaches for Early Diagnosis of Polycystic Ovary Syndrome (PCOS)

Nikshita Chiliveri[1], Harshita Churi[2], Sulaxan Ambade[3], Medha Asurlekar[4]

[1]Department of Artificial Intelligence & Data Science,
K. J. Somaiya Institute of Technology, Sion (E), Mumbai 400022 nikshita.c@somaiya.edu

[2]Department of Artificial Intelligence & Data Science,
K. J. Somaiya Institute of Technology, Sion (E), Mumbai 400022 harshita.c@somaiya.edu

[3]Department of Artificial Intelligence & Data Science,
K. J. Somaiya Institute of Technology, Sion (E), Mumbai 400022 sulaxan.a@somaiya.edu

[4]Department of Artificial Intelligence & Data Science,
K. J. Somaiya Institute of Technology, Sion (E), Mumbai 400022 medha@somaiya.edu

***

Abstract – Polycystic Ovary Syndrome (PCOS) is a common hormonal disorder in women of repro- ductive ages. It is frequently associated with infertility, metabolic abnormalities, and long-term health complications.Due to heterogenous manifestations and the dependence on multiple clinincal factors of this disorder, achieving early and precise diagnosis is difficult. This review provides an overview of machine learning (ML) and deep learning (DL) methods developed for the early detection of PCOS. We examined studies that adopted a range of algorithms including Random Forest, Support Vector Machine, XGBoost, Convolutional Neural Networks, and ensemble methods applied to clinical, biochemical, and ultrasound imaging data. This review describes the important facts about the most frequently utilized datasets, puts an emphasis on key diagnostic markers such as AMH, FSH, LH, BMI, and follicle count, and compares model performance indicators. Particularly, ensemble and stacking approaches showed accuracies above 97%, while explainable AI methods such as SHAP and LIME have improved the trans- parency and clinical interpretability of models. Also, limited diversity of available datasets, inadequate multimodal data fusion, and a lack of extensive validation in real-world clinical environments are some of the long-lasting issues. This paper combines current progress, highlights the existing research gaps, and suggests future pathways for creating robust, interpretable, and clinically applicable PCOS diagnostic tools.

Key Words: Polycystic Ovary Syndrome, Machine Learning, Deep Learning, Early Diagnosis, Explain- able AI, Feature Selection.

1. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is one of the most common endocrine or hormonal disorders in women of reproductive age, with a worldwide prevalence estimated at 6%–20% [13]. It is explained by hyperandrogenism, ovulatory dysfunction, and polycystic ovarian morphology. These factors lead to irregular menstruation, hirsutism, acne, obesity, and insulin resistance. Besides, its impact on reproductive health, PCOS is linked to long-term complications such as Type-2 diabetes, cardiovascular disease, and various psychological health issues.

PCOS diagnosis is conventionally based on the Rotterdam Consensus criteria (2003),which re-

quires at least two of the following three fea- tures must be present: oligo-anovulation, hyperandrogenism, and polycystic ovaries on ultrasound [13]. However, the heterogeneity of PCOS poses strong diagnostic difficulties. Many women face delayed diagnosis due to the wide variation in symptom presentation and inconsistencies in clinical assessment. Early identification is essential for quick management and for mitigating related health risks.

Recent advances in machine learning (ML) and deep learning (DL) have introduced novel strategies for PCOS detection. Models such as Random Forest, Support Vector Machine, and XGBoost show strong performance in classifying PCOS using clinical and biochemical variables [1, 4, 6]. Deep learning models, particularly convolutional neural networks (CNNs), have also showed utility in interpreting ultrasound images [9]. Also the Explainable AI (XAI) approaches, including SHAP [14] and LIME [15], have improved the interpretability of these models, which helps in increasing confidence among healthcare professionals.

In spite of that, several issues persist, such as limited diversity of available datasets, class imbalance, absence of standardized feature sets, and insufficient validation in real-world clinical settings.

1.1 Objectives of This Review

The objective of this review is to thoroughly analyze ML and DL approaches for early PCOS diagnosis by surveying datasets and features used. The review also aims to compare algorithm performances, examine feature selection and XAI techniques, identify research gaps, and propose future directions.

2. RELATED WORK

The application of machine learning and deep learning techniques for PCOS detection has gained significant attention in recent years. Researchers have explored various algorithms, datasets, and feature selection techniques to de- velop accurate and interpretable diagnostic mod- els.

2.1 Machine Learning Approaches

Several studies have applied conventional machine learning techniques to PCOS classification. Zad et al. [1] analyzed Electronic Health Records from Boston Children's Hospital covering 4,500 patients and used Logistic Regression, Random Forest, Gradient Boosting Machine, SVM, and Neural Networks, reporting an AUC of 0.92. Their findings showed that ML models can detect likely PCOS cases even in the absence of explicit diagnostic codes, with elevated testosterone, higher BMI, and menstrual irregularities identified as major predictive factors.

Elmannai et al. [4] introduced a PCOS detec- tion model using Genetic Algorithm for optimized feature selection combined with Random Forest classifier, achieving 98.2% accuracy on a Kaggle dataset of 541 married women. The approach incorporated SHAP for model interpretability, highlighting AMH, FSH, and LH as the most influential features. In a related effort, Khanna et al. [6] designed a three-level stacked ensemble (STACK-3) that integrates multiple classifiers, obtaining 98% accuracy using Mutual Information-based feature selection on the same Kaggle dataset.

Gencer and Gencer [7] evaluated various feature selection strategies, including MARS, Bagging, and Boosting with Multilayer Perceptron. They found that MARS-based selection using 11 features achieved 91.31% accuracy, surpassing models trained on all 41 features. Similarly, Faris and Miften [8] employed a Genetic Algorithm with SVM, reducing the feature set from 42 to 7 while still attaining 90% accuracy with an RBF kernel.

2.2 Deep Learning Approaches

Deep learning methods have also been explored for PCOS detection. Ahmad et al. [3] applied SMOTE-based data balancing with lightweight deep learning models including LSTM, CNN, and hybrid CNN-LSTM architectures. The CNN-based model achieved 96.59% accuracy with minimal training time of 10.02 seconds, demonstrating potential for early-stage detection.

A tri-stage CNN-based wrapper approach was presented by Abouhawwash et al. [12] and achieved a high accuracy of 98.67% on Kaggle data. At the same time, Chelliah et al. [11] compared several algorithms, including ELM, LSTM, and DBN. They found that the Deep Belief Net-

work performed better than the others, with a 97% success rate in all evaluation metrics.

2.3 Image-Based Detection

Divekar and Sonawane [9] focused on PCOS classification from ultrasound images using transfer learning methods. Inception V3 performed the best among all the models tested, which achieved an accuracy of 90.52% and a notably high recall of 97.16%, which is important to ensure that fewer PCOS cases are missed. The study was carried out using the PCOSGen dataset, which contains 4,668 ultrasound images.

2.4 Hybrid and Ensemble Approaches

Recent studies focused on ensemble and stack- ing methods to improve performance of the mod- els. Stacked Learning Framework was proposed in Emara et al. [10] using ADASYN as data balancing method and BORUTA for feature selection, with XGBoost as meta-classifier. The model achieved 97% accuracy. Shaufee et al. [2] combined Particle Swarm Optimization with SVM, achieving 90.18% accuracy through optimized feature selection.

2.5 Research Gaps Identified

According to the surveyed literature, several gaps emerge: (a) heavy reliance on a single Kaggle dataset limits generalizability [4, 6–8, 10–12], (b) most studies lack real-world clinical validation, (c) limited exploration of multimodal approaches combining clinical and imaging data, (d) insufficient focus on diverse populations across different regions, and (e) computational efficiency concerns for real-time clinical deployment [9].

2.6 Key Observations from Literature

From Table 1, several trends emerge. The Kaggle dataset by Kottarathil (541 samples) is the most frequently used benchmark, appearing in 9 of 12 reviewed studies. Random Forest and ensemble/stacking methods consistently achieve the highest accuracies (97-98%) [4, 6]. Feature selection techniques significantly improve model performance while reducing computational complexity [7, 8]. Studies incorporating Explainable AI (SHAP, LIME) provide better clinical interpretability [4, 6, 11]. Image-based approaches using ultrasound data remain underexplored compared to clinical feature-based methods [9].

3. METHODOLOGY

This section outlines the systematic approach adopted for conducting this review, including the literature search strategy, selection criteria, and data extraction process.

3.1 Literature Search Strategy

A comprehensive literature search was conducted across multiple academic databases, including PubMed, IEEE Xplore, ScienceDirect, ResearchGate, Google Scholar, and DOAJ (Directory of Open Access Journals). Combinations of keywords such as “Polycystic Ovary Syndrome,” “PCOS detection,” “PCOS prediction,” “machine learning,” “deep learning,” “artificial intelligence,” “classification,” “diagnosis” and “explainable AI” were used for the search. The search was limited to publications from 2021 to 2025 to capture recent advances in the field.

3.2 Inclusion and Exclusion Criteria

For this review, studies were chosen based on a set of inclusion criteria. These included peer-reviewed journal articles or conference papers that focused on PCOS detection or prediction using ML or DL methods. Only studies that reported clear quantitative performance measures, such as accuracy, precision, recall, F1-score, or AUC, were considered. In addition, the research needed to make use of clinical, biochemical, lifestyle, or imaging data for PCOS classification and be published in the English language.

On the other hand, certain studies were left out of the review. These included works that dealt only with PCOS treatment or management with- out any diagnostic or predictive modeling, as well as review articles, editorials, or opinion pieces that did not present original experimental results. Studies with incomplete methodological details or missing performance metrics were also excluded, along with duplicate publications or research that reused identical datasets and methods.

3.3 Data Extraction Process

From each selected study, key details were collected and organized into a table. This infor-

Table 1: Comparative Analysis of PCOS Detection Studies

Study (Year)	Dataset	Size	Key Features	Models Used	Best Performance	Limitations
Zad et al. (2021)	Boston Children's Hospital EHR	4,500	Age, BMI, testosterone, LH, FSH, insulin	LR, RF, GBM, SVM, NN	AUC: 0.92, Acc: 88%	Single-institution, imbalanced data
Shaufee et al. (2024)	Kaggle (Kapoor)	N/S	42 features (hormonal, lifestyle)	PSO-SVM	Acc: 90.18%	Static dataset, no clinical validation
Ahmad et al. (2024)	PCOS dataset (SMOTE)	N/S	Hormonal features, testosterone	LSTM, CNN, CNN-LSTM	Acc: 96.59% (CNN)	Dataset unclear, no interpretability
Elmannai et al. (2023)	Kaggle (Kottarathil)	541	41 features (BMI, FSH, LH, AMH)	RF, XGBoost, DT, KNN, SVM	Acc: 98.2% (RF)	Small dataset, married women only
Priyadharshini et al. (2024)	Clinical reports	N/S	Age, BMI, HB, RBS, hormonal levels	LR, DT, RF	Acc: 100% (DT)	Overfitting risk, unclear dataset
Khanna et al. (2023)	Kaggle (Kottarathil)	541	43 features (invasive & non-invasive)	STACK-3, DNN, 1D-CNN	Acc: 98% (STACK-3)	XAI incompatibility, regional data
Gencer & Gencer (2023)	Kaggle	541	41 features reduced to 11 (MARS)	MLP with MARS/Bagging	Acc: 91.31% (MARS)	Limited dataset, no external validation
Faris & Miften (2022)	Kaggle (Kottarathil)	541	42 reduced to 7 features	GA-SVM (RBF)	Acc: 90%, Prec: 92%	Small dataset, regional limitation
Divekar & Sonawane (2024)	PCOSGen (Ultrasound)	4,668	Image pixel intensity	Inception V3, ResNet101, ViT	Acc: 90.52%, Rec: 97.16%	Slow inference (~30s/image)
Emara et al. (2025)	Kaggle	541	43 features (LH, FSH, AMH)	Stacked (LR, RF, KNN + XGBoost)	Acc: 97%, Rec: 96%	Small dataset, computationally intensive
Chelliah et al. (2024)	Kaggle (Kottarathil)	541	43 features (SSO, MI, Chi-Square)	ELM, IF, DT, LSTM, DBN	Acc: 97% (DBN)	Single public dataset
Abouhawwash et al. (2023)	Kaggle	541	39 features (tri-stage wrapper)	CNN, MLP, RNN, Bi-LSTM	Acc: 98.67% (CNN)	Single dataset, no multimodal

Legend: LR: Logistic Regression, RF: Random Forest, DT: Decision Tree, SVM: Support Vector Machine, GBM: Gradient Boosting Machine, NN: Neural Network, MLP: Multilayer Perceptron, CNN: Convolutional Neural Network, LSTM: Long Short-Term Memory, DBN: Deep Belief Network, PSO: Particle Swarm Optimization, GA: Genetic Algorithm, MARS: Multivariate Adaptive Regression Splines, N/S: Not Specified

mation included the authors and year of publication, dataset source and size, features, preprocessing steps, the machine learning or deep learning models used, feature selection techniques, performance metrics, key findings, limitations, and future improvements. In total, 12 studies were identified and reviewed.

3.4 Review Framework

The extracted data was arranged in a comparative framework based on four main aspects: (i) datasets and data sources, (ii) feature types and feature selection methods, (iii) classification models and architectures, and (iv) model performance and interpretability. This framework made it easier to compare different studies and to identify common trends as well as gaps in existing research.

4. COMPARATIVE ANALYSIS

This section presents a comprehensive comparative analysis of the discussed studies through the prism of four primary aspects: datasets, features, machine learning and deep learning models, and performance evaluation.

4.1 Datasets and Data Sources

The availability of quality datasets is critical to the development of robust models in the diagnosis of PCOS. Table 2 shows the datasets used as per the studies reviewed.

The most widely used benchmark dataset for PCOS classification studies comes from a Kaggle dataset created by Kottarathil from 10 hospitals in Kerala, India [4, 6–8, 10–12]. The dataset includes 541 rows of female records aged between 20-48 years of reproductive age with 43 features. Although the open availability of the dataset has

Table 2: Summary of Datasets Used in PCOS Detection Studies

Dataset Source	Size	Data Type	Accessibility	Studies Using
Kaggle (Kottarathil)	541 samples	Clinical & Biochemical	Open (CC BY-NC-SA 4.0)	Elmannai et al., Khanna et al., Gencer & Gencer, Faris & Miften, Chelliah et al., Abouhawwash et al., Emara et al.
Boston Children's Hospital EHR	4,500 patients	Electronic Health Records	Restricted	Zad et al.
Kaggle (Kapoor)	Not specified	Clinical features	Open	Shaufee et al.
PCOSGen	4,668 images	Ultrasound Images	Open	Divekar & Sonawane
Clinical Reports	Not specified	Patient observations	Closed	Priyadharshini et al.

been instrumental in advancing PCOS classification studies, there still exists a doubt with the increased dependence on the dataset and the effectiveness of the machine learning model.

The EHR dataset from Boston's Children Hospital used by Zad et al. [1] is the largest available, holding 4,500 patients, but accessibility issues prevent it from aiding in reproducibility or external validation. The second choice in consideration is the PCOSGen dataset, which consists of 4,668 ultrasound image examples, as applied by Divekar and Sonawane. [9]. The issue, however, is that image-based studies of PCOS, as indicated by the provided literature, have not been extensively explored.

4.2 Feature Analysis

Features used in PCOS detection can be categorized into six major groups. Table 3 presents the feature taxonomy identified across studies.

Hormonal and biochemical markers consis- tently emerge as the most discriminative features for PCOS detection. The top predictors identified across multiple studies utilizing explainable AI techniques are AMH (Anti-Müllerian Hormone) and FSH/LH ratio [4, 6, 10]. Physical indica- tions including hair growth, skin darkening, and weight gain also demonstrate strong predictive value [6,8,11], offering non-invasive indicators for preliminary screening.

4.2.1 Feature Selection Methods

Feature selection plays a crucial role in improving model accuracy while reducing computational complexity. Table 4 compares feature selection techniques employed across studies.

The studies show that reducing features from 41-43 to 7-15 can maintain or even improve classification accuracy while significantly reducing model complexity [7,8]. Genetic Algorithm-based selection achieved optimal results in multiple studies [4, 8], successfully identifying the most discriminative feature subsets. Mutual Information combined with ensemble classifiers emerged as the most effective filter-based method [6].

4.3 Machine Learning and Deep Learning Models

Various ML and DL algorithms have been applied for PCOS classification. Table 5 provides a comprehensive comparison of models used across studies.

4.3.1 Model Category Analysis

The reviewed studies broadly fall into three main model categories. Traditional machine learning models, especially Random Forest [4] and XGBoost [4, 10], consistently show strong performance, with reported accuracies ranging from 95–98%. As a result, these approaches are the most commonly used in PCOS classification. Their relatively good interpretability and low computational cost also make them suitable for potential clinical use.

Deep learning models have shown encouraging results, particularly in image-based PCOS classification. CNN-based approaches achieved accuracies of up to 98.67% on clinical datasets [12] and demonstrated high recall values of 97.16% when applied to ultrasound images [9]. However, when applied to small tabular datasets, deep learn- ing models often performed worse than ensemble-

Table 3: Feature Taxonomy for PCOS Detection

Category	Features	Frequency
Demographic	Age, weight, height, BMI, blood group, marital status	High (100%)
Hormonal/ Biochemical	FSH, LH, FSH/LH ratio, AMH, TSH, PRL, PRG, testosterone, insulin, Vitamin D3, β-HCG	High (100%)
Reproductive Health	Menstrual cycle regularity, cycle length, pregnancy history, number of abortions, follicle count (L/R), follicle size, endometrium thickness	High (92%)
Physical/ Clinical Signs	Hirsutism, hair growth, skin darkening, acne/pimples, hair loss, weight gain, waist-hip ratio	High (83%)
Vital Signs	Blood pressure (systolic/diastolic), pulse rate, respiration rate, hemoglobin	Medium (67%)
Lifestyle Factors	Fast food consumption, regular exercise, sleep patterns	Medium (58%)
Imaging Features	Ultrasound image pixel intensity, ovarian morphology	Low (17%)

based methods [6], which may be attributed to limited sample sizes.

The highest overall performance is achieved by Ensemble and Stacking approaches. The multi-level STACK-3 model by Khanna et al. [6] and the stacked framework by Emara et al. [10] both achieved 97-98% accuracy.

4.4 Performance Evaluation

Table 8 presents a comprehensive performance comparison across all reviewed studies.

4.4.1 Analysis of Results

Several important findings were revealed during the performance analysis. Accuracy alone is insufficient as a metric, particularly given the class imbalance in PCOS datasets. Studies that ap- plied data balancing techniques such as SMOTE, ADASYN, and Borderline-SMOTE generally reported more reliable performance, with balance between precision and recall values [3, 6, 10].

More importantly , Recall is particularly criti- cal for PCOS detection as false negatives (missed diagnoses) can lead to delayed treatment and complications. Divekar and Sonawane's Inception V3 model [9] achieved the highest recall (97.16%), making it valuable for screening applications that prioritizes sensitivity.

4.5 Explainable AI Integration

The use of Explainable AI (XAI) techniques has become an important trend in recent PCOS detection studies, as summarized in Table 9.

The integration of XAI has been identified as a potential solution to one of the major barriers to the adoption of medical AI, which is transparency in decisions [14, 15]. ome studies found that features emphasized by XAI methods, such as SHAP values and LIME, strongly correlate with traditional medical knowledge, including the well-known Rotterdam criteria [13]. However, Khanna et al. [6] found XAI methods had difficulty assisting with complex stacking architectures.

4.6 Summary of Comparative Analysis

As seen from this comparative analysis, it is found that the maximum performance in the detection of PCOS cases is achieved with ensemble and stacking techniques, combined with feature selection and balancing techniques [6, 10, 12]. Among all single-classification techniques, the Random Forest classifier is again seen to be the most reliable [4], with CNN techniques having high possibilities as image diagnosis techniques [9, 12]. The incorporation of XAI techniques further improves the clinical relevance of these models by enhancing the models' transparency [4, 6, 11], however, one drawback of ensemble models is their difficulty in interpretation. As seen from this review, the limitation associated with all the models is the reliance on small and restricted datasets.

5. DISCUSSION

This section brings together the results of the comparative analysis, highlights existing research gaps, and discusses the main challenges in current PCOS diagnostic systems.

Table 4: Feature Selection Methods in PCOS Detection

Method	Type	Study	Original	Selected	Impact on Accuracy
Genetic Algorithm	Wrapper/ Metaheuristic	Elmannai et al.	41	Optimized	Improved to 98.2%
Genetic Algorithm	Wrapper/ Metaheuristic	Faris & Miften	42	7	Maintained 90%
MARS	Statistical	Gencer & Gencer	41	11	Improved to 91.31%
Bagging	Ensemble	Gencer & Gencer	41	10	88.72%
Boosting	Ensemble	Gencer & Gencer	41	4	87.61%
Mutual Information	Filter	Khanna et al.	43	15	Best with STACK-3 (98%)
Harris Hawk Optimization	Metaheuristic	Khanna et al.	43	15	Comparable
PSO	Metaheuristic	Shaufee et al.	42	Optimized	90.18%
BORUTA	Wrapper	Emara et al.	43	Subset	97%
Tri-stage Wrapper	Wrapper	Abouhawwash et al.	39	Optimized	98.67%
Chi-Square Test	Filter	Chelliah et al.	43	Subset	97%

5.1 Key Findings

The review of the recent studies shows clear patterns. Ensemble and Stacking methods applied in the studies perform better than individual sin- gle classifiers, achieving the accuracy between 97- 98% [6, 10, 12]. On the other hand , in the single classifiers , Random Forest proves to be the most reliable one. [4]. Also, the Deep Learning models show better results in the Image-Based PCOS data and it's detection, [9], but they do not perform well on small and tabular datasets. In such cases , the traditional Machine Learning Models show better approach on the tabular data. [6].

Secondly , we observe the role of feature selec- tion in various studies. It shows that the reduction in input features from 41 - 43 features to 7-15 features, does not harm the performance of the model. In fact in some cases, it improves the accuracy of the model [7, 8]. We also analyzed in the studies that, hormonal factors such as AMH, FSH and LH as well as physical indicators like follicle count, BMI and hair growth are proven to be the most discriminative feature in these studies [4, 6, 10]. These factors align with the Rotterdam diagnostic criteria [13].

Our another observation from analysis, is that the class imbalance in the PCOS data is a major challenge. in order to mitigate this challenge various data balancing methods were used in several studies. The most used techniques were SMOTE and ADASYN, showing more stable results in terms of balancing the precision-recall trade-off [3, 6, 10]. In order to improve the interpretabil- ity of the model and its prediction, XAI tech- niques (SHAP, LIME) were integrated. This helps in aligning them with known medical knowledge. [4, 6, 11, 14, 15].

5.2 Research Gaps

Although the latest studies have reported considerable progress, several issues remain unanswered. The most notable concern is the fact that most studies rely heavily on the Kaggle dataset used in almost 70% of the work. [4, 6–8, 10–12]. The dataset has only 541 samples. Naturally, the effectiveness of the models in this study in generalizing beyond the dataset remains in question. Another issue is the fact that the data is from Kerala in India and comprises only married women. [6, 8].

Another limitation is that most studies rely on a single data modality. Existing approaches typically use either clinical data or ultrasound images in isolation, even though PCOS diagnosis in practice involves information from multiple sources. Effective multimodal integration is largely absent from current research. Furthermore, most studies are based on retrospectively curated datasets and do not include prospective clinical validation [1].

Table 5: Traditional Machine Learning Models for PCOS Detection

Model	Studies Using	Best Acc.	Strengths	Limitations
Random Forest	Zad et al., Elmannai et al., Khanna et al., Priyadharshini et al., Emara et al.	98.2%	High accuracy, handles imbalanced data, feature importance	Less interpretable than single trees
SVM	Zad et al., Shaufee et al., Elmannai et al., Khanna et al., Faris & Miften	93.2%	Effective for high-dimensional data, kernel flexibility	Sensitive to parameter tuning
XGBoost	Elmannai et al., Khanna et al., Emara et al.	97.5%	Handles missing data, regularization	Prone to overfitting on small datasets
Decision Tree	Elmannai et al., Priyadharshini et al., Khanna et al., Chelliah et al.	100%*	High interpretability, fast inference	Overfitting risk (*likely overfit)
KNN	Elmannai et al., Khanna et al., Emara et al.	91.5%	Simple, no training phase	Sensitive to scaling, slow inference
Logistic Regression	Zad et al., Khanna et al., Emara et al.	88%	Interpretable, probabilistic output	Limited for non-linear relationships
MLP	Gencer & Gencer, Abouhawwash et al.	91.31%	Captures non-linear patterns	Requires more data

As a result, there is little evidence of real-world performance. Longitudinal analyses that track disease progression over time are also missing from the literature.

5.3 Challenges

One of the biggest challenges in PCOS research is still data availability. Building large and diverse datasets is difficult, mainly because of privacy regulations and institutional barriers. Although several studies utilize artificial oversampling techniques, these methods by themselves seem to be inadequate to deal with this imbalance issue effectively. Another issue with implementing these models is the availability of several highly effective features, which necessitate costly and invasive medical procedures, thus restricting their implementation in underprivileged areas [6, 8].

Moreover, it is clear that there is also a trade- off between model accuracy and clinical usabil- ity. Models that have the best performance may be quite complex, and the prediction is difficult for the clinician to understand [6]. Further constraints are added with the use of deep learning models, which require considerable computational powers and inference time, which may not be possible for clinical use [9]. Going beyond the aforementioned issues, there is still the topic of regulatory approval, clinical responsibility, and the ethical problems of bias in algorithms.

6. FUTURE DIRECTIONS

This section discusses the future research direction based on the gaps identified from the existing research in PCOS.

6.1 Dataset Development

Looking into the future, more effort appears to be required to build larger datasets for the future research into PCOS. This might need collaboration between more than one hospital and research institutions across different regions, rather than relying on data collected from just a single source. Looking into the future, datasets considered might be those collected from women belonging to different age groups, ethnicities, and demographic settings to ensure that they are not restricted to a certain type of population only. It is important that data is collected using universal protocols, which would be beneficial for ensuring more consistency within the data and making the results more comparable and verifiable.

6.2 Multimodal Learning Approaches

The fusion of different types of information, such as clinical characteristics, hormonal test results,

Table 6: Deep Learning Models for PCOS Detection

Model	Studies Using	Best Acc.	Strengths	Limitations
CNN	Ahmad et al., Abouhawwash et al., Divekar & Sonawane	96.59% - 98.67%	Excellent for image data, automatic feature extraction	Requires large data, high computation
LSTM	Ahmad et al., Chelliah et al.	92.04%	Captures sequential patterns	Limited benefit for tabular data
CNN-LSTM	Ahmad et al.	94.31%	Combines spatial and temporal features	Complex architecture
DBN	Chelliah et al.	97%	Effective for unsupervised pretraining	Training complexity
1D-CNN	Khanna et al.	90%	Efficient for 1D clinical data	Underperformed vs. ensemble
DNN	Khanna et al.	93.85%	Flexible architecture	Requires hyperparameter tuning
Inception V3	Divekar & Sonawane	90.52%	Pretrained, multi-scale features	Slow inference (~30s/image)

Table 7: Ensemble and Stacking Models for PCOS Detection

Model	Studies Using	Best Acc.	Configuration
STACK-3 (Multi-level)	Khanna et al.	98%	STACK-1 + STACK-2 combined
Stacked (ADASYN + BORUTA)	Emara et al.	97%	LR, RF, KNN + XGBoost meta-classifier
AdaBoost	Khanna et al.	~95%	Boosted weak learners
Extra Trees	Khanna et al.	~96%	Randomized decision trees
PSO-SVM	Shaufee et al.	90.18%	PSO-optimized SVM parameters
GA-SVM	Faris & Miften	90%	GA feature selection + RBF SVM

ultrasound test results, and lifestyle, in one model is an encouraging trend in the future of the study of PCOS [1, 4, 6, 9]. The use of multimodal fusion networks can assist in the fusion of various pieces of information from different types, causing the comprehensive and precise diagnosis of the disease. In addition, the process is more realistic because, in clinical practice, physicians do not solely rely on a certain type of evidence in diagnosing a patient with PCOS.

6.3 Non-Invasive Screening Tools

In practical situations, much value can be placed in emphasizing PCOS models that rely on non-invasive information. Information like menstrual patterns, weight, and other evident clinical symptoms are already discussed during the early stages of clinical consultation and do not require laboratory investigations. [6, 8]. Carefully developed PCOS models based on such information can aid the creation of simple screening tools through online or mobile interfaces. It could aid the early identification from a health point of view, though it does not constitute a real diagnosis, particularly for women living in resource-constrained regions.

6.4 Longitudinal Analysis

Longitudinal data can also potentially show how PCOS develops and how it changes over time; therefore, this is another area to be explored in future studies. In this regard, it may prove beneficial to try to link diagnostic models with wearable technology and menstrual tracking apps. This could potentially allow for continuous monitoring to assess PCOS more realistically over time.

6.5 Advanced Explainable AI

There is still a clear need to develop XAI methods that work effectively with complex ensemble models [6]. While these models often achieve strong performance, their decision-making processes are difficult to interpret in clinical settings. Focus- ing on inherently interpretable models that can maintain high accuracy while offering clear and transparent decision pathways would help improve

Table 8: Performance Metrics Comparison Across Studies

Study	Best Model	Acc.	Prec.	Rec.	F1	AUC	Data Balancing
Zad et al. (2021)	Neural Network	88%	87%	84%	-	0.92	Not specified
Shaufee et al. (2024)	PSO-SVM	90.18%	-	-	-	-	None
Ahmad et al. (2024)	CNN	96.59%	-	-	-	0.966	SMOTE
Elmannai et al. (2023)	Random Forest	98.2%	97.8%	98.3%	98%	-	Not specified
Priyadharshini et al. (2024)	Decision Tree	100%*	-	-	-	-	None
Khanna et al. (2023)	STACK-3	98%	97%	98%	98%	1.00	Borderline-SMOTE
Gencer & Gencer (2023)	MLP (MARS)	91.31%	-	-	-	-	None
Faris & Miften (2022)	GA-SVM	90%	92%	75.7%	83%	-	None
Divekar & Sonawane (2024)	Inception V3	90.52%	90.01%	97.16%	93.45%	-	None
Emara et al. (2025)	Stacked + XGBoost	97%	-	96%	-	-	ADASYN
Chelliah et al. (2024)	DBN	97%	97%	97%	97%	-	Not specified
Abouhawwash et al. (2023)	Tri-stage + CNN	98.67%	97%	89%	-	-	Not specified

*Likely indicates overfitting due to small dataset

trust and acceptance among clinicians [14, 15].

6.6 Clinical Validation and Deployment

Before AI models can be used in real clinical settings, they need to be tested through prospective clinical trials that reflect everyday healthcare conditions. Such trials help ensure that the models work reliably with real patient data and fit naturally into clinical workflows. In addition, us- ing lightweight and efficient model designs can make these systems easier to deploy and main- tain. Integrating AI tools with Electronic Health Records [1] would further support smooth adoption by allowing clinicians to access predictions directly within the systems they already use.

7. CONCLUSION

This review examined recent machine learning and deep learning approaches used for the early diagnosis of Polycystic Ovary Syndrome. A total of 12 relevant studies published between 2021 and 2025 were analyzed, with a focus on the datasets used, selected features, applied algorithms, and reported performance measures in PCOS detection models.

Overall, the findings show that ensemble and stacking techniques provide the highest diagnostic accuracy, typically in the range of 97–98% [6, 10, 12]. Among individual models, Random Forest consistently is considered to be the most reliable classifier [4]. Feature selection methods, especially Genetic Algorithm [4, 8] and Mutual Information [6], play an important role in improving performance while keeping models less complex. Across studies, hormonal markers such as AMH, FSH, and LH, along with physical indicators including follicle count, BMI, and hair growth, are repeatedly identified as key predictors [4, 6, 8, 10, 11]. These findings are consistent with the Rotterdam diagnostic criteria [13]. In addition, data balancing techniques like SMOTE and ADASYN are shown to be essential for managing class imbalance [3, 6, 10]. The use of Explainable AI methods, particularly SHAP and LIME, further improves clinical interpretability and builds trust in model predictions [4, 6, 11, 14, 15].

Table 9: Explainable AI Methods in PCOS Detection Studies

Study	XAI Methods	Key Interpretable Findings	Clinical Relevance
Elmannai et al. (2023)	SHAP	AMH, FSH, LH identified as top predictors	Aligns with Rotterdam criteria
Khanna et al. (2023)	SHAP, LIME, ELI5, Qlattice, Feature Importance	Follicle count, hair growth, weight gain, skin darkening as key features	Supports both invasive and non-invasive screening
Chelliah et al. (2024)	SHAP, LIME	Follicle count, skin darkening identified	Provides transparent decision support
Divekar & Sonawane (2024)	LIME	Visual explanations for ultrasound classification	Highlights image regions influencing prediction

Despite these advances, several limitations remain. Many studies rely heavily on a single small dataset containing only 541 samples, which raises concerns about how well the models generalize to broader populations [4, 6–8, 10–12]. Most existing work focuses on a single data type and lacks effective multimodal integration, prospective clinical validation, and longitudinal analysis. Practical issues such as computational efficiency [9] and limited model interpretability [6] also continue to restrict real-world deployment.

Future research should therefore prioritize the development of larger and more diverse multicenter datasets, better integration of multiple data sources, and the design of non-invasive screening tools. Conducting prospective clinical studies will be essential to validate model performance in real healthcare settings. Further progress in explainable AI for complex models and optimization for low-resource or edge deployment will also be important for turning research into practice.

In summary, machine learning and deep learn- ing methods show strong potential for early PCOS detection and could support earlier intervention and improved health outcomes. Addressing the current limitations will be key to developing reliable, interpretable, and clinically usable PCOS diagnostic systems.

ACKNOWLEDGEMENT

The authors would like to thank the Department of Artificial Intelligence & Data Science at K. J. Somaiya Institute of Technology, Mumbai, for their support, guidance, and resources throughout the course of this research.

References

[1] Zad, Z., Jiang, V.S., Wolf, A.T., Wang, T., Cheng, J., Paschalidis, I.C., Mahalingaiah, S.: Predicting Polycystic Ovary Syndrome with Machine Learning Models Using Electronic Health Records. PMC (2021). https://pmc.ncbi.nlm.nih.gov/articles/PMC10866556/

[2] Shaufee, L.H., Jantan, H., Bahrin, U.F.M.: Polycystic Ovary Syndrome (PCOS) Prediction System Using PSO-SVM. DOAJ (2024). https://doaj.org/article/d5349f8874134e5a88e2384f851485e3

[3] Ahmad, R., Maghrabi, L.A., Khaja, I.A., Maghrabi, L.A., Ahmad, M.: SMOTE- Based Automated PCOS Prediction Using Lightweight Deep Learning Models. PubMed (2024). https://pubmed. ncbi.nlm.nih.gov/39410629/

[4] Elmannai, H., El-Rashidy, N., Mashal, I., Alohali, M.A., Farag, S., El-Sappagh, S., Saleh, H.: Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Ex- plainable Artificial Intelligence. Diagnos- tics (2023). https://pmc.ncbi.nlm.nih.gov/articles/PMC10137609/

[5] Priyadharshini, M., Srimathi, A., Sanjay, C., Ramprakash, K.: PCOS Disease Prediction Using Machine Learn- ing Algorithm. ResearchGate (2024). https://www.researchgate.net/publication/379192773

[6] Khanna, V.V., Chadaga, K., Sampathila, N., Prabhu, S., Bhandage, V., Hegde, G.K.:

A Distinctive Explainable Machine Learn- ing Framework for Detection of Polycys- tic Ovary Syndrome. Applied System In- novation, MDPI (2023). https://www. mdpi.com/2571-5577/6/2/32

[7] Gencer, K., Gencer, G.: Machine Learning Assisted Decision Mak- ing in Patients with Polycystic Ovary Syndrome. ResearchGate (2023). https://www.researchgate. net/publication/370224738

[8] Faris, N.N., Miften, F.S.: Detection of PCOS Based on Genetic Algorithm Cou- pled with SVM. ResearchGate (2022). https://www.researchgate.net/ publication/381104998

[9] Divekar, A., Sonawane, A.: Leveraging AI for Automatic Classification of PCOS Using Ultrasound Imaging. arXiv (2024). https: //arxiv.org/pdf/2501.01984

[10] Emara, H.M., El-Shafai, W., et al.: A Stacked Learning Framework for Accurate Classification of Polycystic Ovary Syn- drome with Advanced Data Balancing and Feature Selection Techniques. PMC (2025). https://pmc.ncbi.nlm.nih.gov/ articles/PMC12088954/

[11] Chelliah, B.J., Gahra, S.K., Senthil- selvi, A.: Enhancing PCOS Predic- tion Using Machine Learning and Explainable AI. ResearchGate (2024). https://www.researchgate.net/ publication/388789322

[12] Abouhawwash, M., Sridevi, S., et al.: Automatic Diagnosis of Polycystic Ovarian Syndrome Using Wrapper Methodol- ogy with Deep Learning Techniques. Computer Systems Science and Engi- neering, TechScience Press (2023). https://www.techscience.com/ csse/v47n1/53012/html

[13] Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group: Re- vised 2003 Consensus on Diagnostic Criteria and Long-Term Health Risks Related to Polycystic Ovary Syndrome. Fertility and Sterility, Vol. 81, No. 1 (2004) 19–25

[14] Lundberg, S.M., Lee, S.I.: A Unified Ap- proach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, Vol. 30 (2017) 4765–4774

[15] Ribeiro, M.T., Singh, S., Guestrin, C.: Why Should I Trust You? Explaining the Pre- dictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Con- ference on Knowledge Discovery and Data Mining (2016) 1135–1144