International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Explainable Machine Learning Approaches for Accurate Prediction of Polycystic Ovary Syndrome: A Review

Astha Meena[1], Dr. Devendra Singh Rathore[2], Dr. Vivek Richhariya[3]
[1] Research Scholar, Department of CSE, Lakshmi Narain College of Technology, Bhopal, India
[2] Associate Professor, Department of CSE, Lakshmi Narain College of Technology, Bhopal, India
[3] Professor, Department of CSE, Lakshmi Narain College of Technology, Bhopal, India

Abstract:
Polycystic Ovary Syndrome (PCOS) is a widely recognized endocrine disorder affecting women of reproductive age, causing severe complications, such as infertility, metabolic disturbances, and psychological distress. Early detection and accurate diagnosis are still challenging tasks because of heterogeneous symptoms with no specific diagnostic methods. This paper elaborated the design-development of SmartScanPCOS, a feature-based framework creatively inspired by advanced Machine Learning (ML) techniques conjugated with Explainable Artificial Intelligence (XAI) with simultaneous influence on both prediction and interpretation of a PCOS diagnosis. The paper discovered from notes critically the progress in machine-learning (ML) prediction for PCOS, basing the prediction on analytical biomarkers of hormonal levels, metabolic signal, life course markers, which generally involve diverse disease-like features; further predictive modeling takes into account medical imaging/engineering data to assess improved performance for interpretation. Several classification-based algorithms with some other major models (also compared to general linear statistical methods) within several experimental protocols plan for an effective technique; computations had to be in random forests, support vector machines, and neural networks, based on predictive comparison: Thus, some XAI methods, including LIME and SHAP, might be confirmed as interpreting the explanation by making the most influential attribute in the model decide. This is how a model provides healthcare practitioners and other services with confidence to make better decisions. The proposal, therefore, corresponds well to the issues of feature selection, model optimization, and interpretability with high class accuracy. The conclusion reached after literature review opportunities is the existence and allied collaboration of hybrid and ensemble predictors with explanation capabilities. The finding suggested possible inroads in providing some means for bridging the gap in further-processed data-driven elucidation vis-à-vis full-scale clinical considerations of PCOS diagnosis.

Keywords: PCOS Prediction, Machine Learning, Explainable AI, Feature Selection, Healthcare Analytics, Smart Diagnosis.

1. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a prevalent guideline of endocrine and metabolic disorders affecting women of childbearing age in the modern era. It displays a series of symptoms that combine to suggest such a diagnosis, irregular menses cycles, hyperandrogenemia, polycystic ovarian morphology, insulin resistance, and metabolic abnormalities. Being widespread, PCOS is sadly often underdiagnosed by clinicians being unsure how to manage such an informally designated and ill-defined group of sufferers with a variety of symptom clusters. Actually, there remains no golden embryo to confirm the diagnosis of PCOS-can it ever be produced- as the syndrome is heterogeneous with respect to clinical presentation and relevant diagnostic tests are most uncertain. Traditional diagnostic criteria, such as the Rotterdam criteria, require many laboratory and imaging parameters that differ from patient to patient. This inconsistency introduces a great deal of diagnostic doubt, hence contributing to the delay of treatment and elevation of complications in the long run, such as increasing the risk of type 2 diabetes, as well as cardiovascular diseases and infertility. With the high advance of technologies in digital healthcare as well as data-oriented methodologies in recent years, these areas have hugely opened new ways to outlook on disease diagnosis and management [1]. Among these Machine Learning (ML) came forth as hugely powerful to discern hidden links in complex, high-dimensional medical data. ML algorithms work in such a way that they might burn large amounts of structured and unstructured information (ranging from clinical records, hormonal profile, lifestyle, imaging data, among others) to generate highly accurate predictive models. These models were revealed to be of immense assistance to healthcare professionals in the early detection of PCOS and the risk assessment, thereby ensuring timely intervention and individualized remedy strategies. A cardinal criterion for predilection of PCOS through generation of machine learning is the finding and selection of those features that materially contribute to this disorder. As stated above, the significant approaches aimed at feature extraction try to derive insights from a considerable variety of datasets that contain BMI, insulin levels, follicle count, menstrual irregularities, and certain biochemical markers. Good feature selection primarily enhances model performance. More often than not, good feature selection also lowers the COST or computational complexity, thereby possibly improving generalization. Several methodologies such as filter methods, wrapper methods, and embedded methods have been chronicled in the annals of literature for obtaining best-suited feature subsets for our predictory immodest. It has been a practice for the past few years that Machine Learning models like Support Vector Machines, Random Forest, Decision Trees, and Artificial Neural Networks have been very efficient in diagnosing PCOS. With that annotated, the challenge still remains whether these models do or do offer transparency. Most prominent among the high-performing models, especially the deep learning models, is their functioning as black boxes they give predictions, but how they are making those decisions remained unaccounted for. These are characteristics of the medical domain where clinical decisions directly impact patient outcomes so explanation and trustworthiness should become held in high significance [2]. This led to the increase in the number of people adopting the Explainable Artificial Intelligence (XAI) approach, which aims to improve transparency, interpretability, and accountability of Machine Learning models. Powerful Interpretable Model-blind Explanation methods (LIME) and Shapley Additive Explanations (SHAP) were developed with sufficient awareness to the need of explaining model behavior by highlighting the contribution of each feature towards a certain output. The described methods let clinicians see how the model outputs at the time, thus supporting intelligent decision-making and trust in the AI-interpreted diagnostic system. In case of PCOS, XAI could be useful in identifying unique and hidden risk factors or patterns that may not be recognized in output from a traditional analysis. Such interpretability assists at bridging the gap between artificial intelligence and the clinic. Figure 1 illustrates the structural differences between a normal ovary and a polycystic ovary, highlighting the presence of multiple cysts in PCOS.

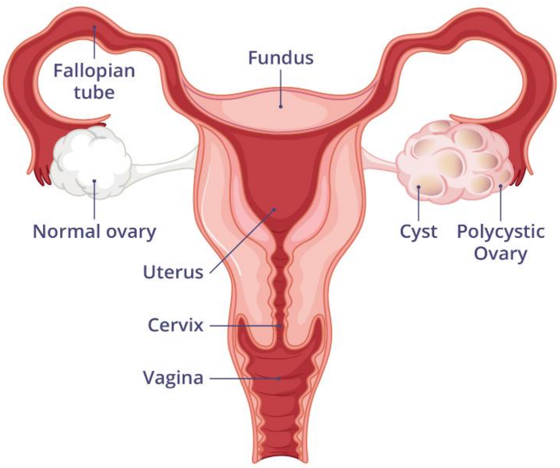

Figure 1: Anatomical Representation of Normal and Polycystic Ovary [2]

The capital achievement for the diagnosis of PCOS is this integration of intelligent methods such as ML and XAI to describe data characteristics into a single framework. The approach aims at some significant improvements related to prediction of diseases while ensuring the concept of interpretability for the decisions. The conceptual framework of SmartScanPCOS distills these ideas into one mirror and presents advanced data analytics, advanced feature selection techniques and explainable models using the very multiprocessing framework. The SmartScanPCOS project offers a major diagnostic support system that holds up the claims of efficiency and reliability with respect to the diverse datasets and modern methodologies. This project intends to clear grounds for the incorporation of early detection, risk stratification, and personalized management functions within a single system for PCOS. This review paper is dedicated to analyzing the present scenario of ML and XAI applications in the prediction of PCOS. It systematically explores various methodologies, datasets, feature engineering techniques, and model evaluation strategies that have been employed by earlier studies.

Additionally, it points out the advantages and limitations of existing approaches, underlining the pressing need for interpretable and clinically interesting models. Relevantly, by comparing methodologies and results of the state of the art, the subject matter gain insights into emerging trends, research deficits, and future opportunities in this area [3].

Thus, PCOS increasingly poses different problems and the presented diagnostic measures are in need of a state-of-the-art empirical technological solution. Teaching machines that work with Explainable AI provide a new opportunity to make more transparent and precise predictions for diagnosing PCOS. The focus on these methods are intended to be feature driven and interpretable, contributing to the development of next-generation technological healthcare systems.

2. OVERVIEW OF POLYCYSTIC OVARY SYNDROME (PCOS)

Polycystic ovary syndrome (PCOS) is an endocrine and metabolic disorder of multifactorial origin that affects 5-10% of women of reproductive age. It is increasingly acknowledged as one of the causes of infertility and is linked with many reproductive, metabolic and psychological disorders. Its inherent complexity lies in its multisymptomatic and fluctuating phenotype, which varies greatly among individuals. These variations represent a fairly remote chance of accurate diagnosis of PCOS and appropriate commening of therapeutic measures. This necessitates a rather deep understanding of clinical or biological aspects of the syndrome. PCOS is characterized by the combination of these three main symptoms - ovulatory dysfunction, hyperandrogenism, and polycystic ovarian morphology [4]. Ovulatory dysfunction takes place in the form of altered or absent periods and can cause significant fertility issues in some women. Hyperandrogenism is marked by male hormone levels, leading to the katagenic pattern of hirsutism, acne, and androgenic alopecia. Polycystic ovarian morphology is pictured on the ultrasound, where there are multiple small follicles present peripherally on the ovary. According to some widely acceptable diagnostic frameworks, the clinical diagnosis of the syndrome requires presence of two out of these three objective findings as, for instance, in the Rotterdam criteria. All patients with the polycystic ovarian syndrome do not necessarily exhibit the same combination of symptoms that otherwise foster diagnostic confusion. PCOS also called polycystic ovarian syndrome is a common reproductive endocrinopathic disorder. The etiology of polycystic ovarian syndrome is complex and remains uncertain and depends on a vast interaction among genetic, hormonal, environmental, and lifestyle factors. Although genetic predisposition plays a major role in PCOS development, they tend to occur within families [5]. Hormonal imbalances are also key contributors to the pathophysiology of the disease, particularly insulin and androgens. In a large number of patients with PCOS, insulin resistance is an important observation, in which insulin fails to stimulate the body's cells effectively. As a phenomenon, there is compensatory hyperinsulinemia that then leads to the ovaries producing more androgens to disrupt the normal follicular development in ovulation. It's a beautiful trilogy. The etiology of PCOS depends heavily on the manifestation of lifestyle elements, especially diet, a sedentary lifestyle, and obesity. Obesity, while contributing to an increase in insulin resistance and hormonal imbalances, augments any of the syndrome's features and enhances the comradeship of associated metabolic disorders. Nevertheless, it is important to stress that PCOS does affect normal-weight females-therefore, obesity could not be a condition but only an auxiliary problem. Additionally, some researchers have asserted environmental strategies like exposure to endocrine-disrupting chemicals as possible contributors; however, more in-depth investigations are required to establish their exact idiosyncratic role.

A profound manifestation of polycystic ovary syndrome, which is beyond reproduction, is a distribution of metabolic complications and psychological aspects. Due to insulin resistance and metabolic disturbances, women with PCOS are at an increased risk of acquiring type 2 diabetes, dyslipidemia, hypertension, and cardiovascular diseases. In addition, patients commonly experience mood disorders such as anxiety and depression, which in turn affect their quality of life. This alone certainly advocates for a holistic form of PCOS management involving both physical and psychological health. [6]

PCOS Diagnosing remains a tough part without a single definitive test and thanks to heterogeneous presentations. Diagnostic practices currently involve historical examinations, the physical examination, laboratory tests, and the demonstration of imaging studies. The hormone assays measure androgen along with insulin levels and LH/FSH levels while ultrasound imaging defines. "Therefore, timely diagnosis of the syndrome and sex hormone secretion in order to begin treatment has become particularly delicate in resource-poor settings." Correct management options for PCOS usually take the individual's symptoms and pregnancy considerations into account. The recommended treatment measures are in three categories: lifestyle alterations, pharma-medical treatment, and assisted reproductive treatments. Lifestyle interventions--e.g., curbing weight, better nutrition, and physical fitness--are common with a combined practice amongst overweight people. The pharmacological options would include then the employment of hormonal contraceptive therapy which helps menstruation problems, insulin sensitizing agents which are used with metformin, and anti-androgens that take care of hyperandrogenism symptoms [7]. With the multiple treatment options for the long-term management of PCOS, its chronic nature and the need for monitoring during the course of treatment still pose a challenge to proper management. Moreover, the lack of awareness and therefore the delayed diagnosis result in many untreated or poorly treated cases which would hence become increasingly prone to complications. For this reason, the early detection and accurate diagnosis serve both to improve patient outcomes and reduce the disease burden. Various technological innovations on the computational side have made great feats in generating data that can be interesting to know about the diagnosis and management of polycystic ovary syndrome (PCOS). Data water algorithms placed in a machine learning process that contains some knowledge-theoretic foundations to compute complex interactions between many variables and explore patterns that may go unnoticed without the 'advanced methods' are some examples of data-driven approaches. On the condition of adequate availability of clinical, biochemical, and lifestyle data, these approaches quite powerfully could derive the best possible prediction in accordance with the peculiar patient profile for an effective and truly personalized patient management. Consequently, these data provide an assistant tool for distinguishing how these many factors apply to a given patient case, truly supporting the physician in actual patient management-associated decision making [8].

Thus, PCOS is a complex, multifactorial affliction whose significance toward women's health is beyond contest. Diverse presentations, hormonal and metabolic alterations causing derangements, and correlated comorbidities complicate its diagnosis and management. Consequently, a detailed comprehension of its pathophysiology, risk factors, and clinical features provides a foundation for proper implementation of diagnostic and therapeutic techniques. Advanced techniques involving analytical tools, as well as a multidisciplinary approach, will continue to prove invaluable as research progresses in the arena of PCOS and, through concerted efforts, improve the standards of care for this affected population.

3. ROLE OF MACHINE LEARNING IN HEALTHCARE DIAGNOSIS

Today, machine learning (ML), a sub-section of Artificial intelligence, is being labeled as a revolutionary technology in the healthcare industry, due to diagnosing diseases and clinical services. As medical data from electronic health records (EHRs), diagnostic medical imaging, wearable devices, and laboratory tests skyrocket, classical methods of data analysis often come to grief in the processing and interpretation of data sets that appear massive and complex. An ML approach performs the grim task of automated learning for acquiring the former hidden patterns. It generates predictive insights with optimum efficiency and accuracy. Hence, ML has become a critical modality within the augmentation of diagnostic precision, errorless work by humans, and support for personal medication. Working with large dimensional and heterogeneous data is one of the primary advantages of machine learning in healthcare diagnosis. Medical datasets invariably consist of a mix of structured data (like patient demographics, laboratory values) and unstructured data (like clinical notes, medical images) [9]. Therefore, machine-learning algorithms can utilize these diverse sources of data to create a comprehensive model that captures complex relationships between the above variables. For instance, in the context of the disease prediction tasks, a machine learning model will have the potential to analyze multiple risks, including genetic predisposition, lifestyle habits, and physiological measures, thereby providing a holistic picture of the patient's condition.

The one who is currently applying cognitive science tools for describing thoughts and experiences is the one. While these methods prove quite useful in accessing some areas from a machine learning perspective, strategies like targeting user-centered therapy should be taken up for optimism, and steps of validating these guitar-oriented AI models can be suggested.

Unsupervised learning, on the other hand, detects the hidden structure or pattern behind unrated data. Clustering and dimensionality reduction are ways of identifying disease subtypes, patient stratification, and anomaly detection. Such an approach in clinical settings can play an important role in putting similar patient characteristics into groups that have been truly helped by more substantial treatment strategies. As an internal mechanism, reinforcement learning is less utilized. However, research has shown its potential in the optimum enhancement of treatment plans and decision-making processes precisely by learning dynamically from interactions with the environment. An important role of ML in the diagnosis of diseases is the capacity to detect them at an early stage. At times, advanced diagnosis is very important for effective treatment and post-surgery prognostics, especially in cases of chronic and multilateral disease conditions. A machine learning model can pick up and understand subtle patterns and correlations that human clinicians may miss, thereby detecting a sick condition as the very early stage. For example, if, using a predictive model, historical and real-time data indicates a very high chance of developing diabetes, cardiovascular diseases, PCOS, or endocrine disorders, treatment can be administered early to prevent common, unpleasant and debilitating consequences [10].

Feature dimensionality and feature selection have a significant influence on the success of machine learning-based diagnostic systems. Variables that are highly related strengthen the predictive potential and considerably decrease complexity. One frequently utilized feature selection method is the use of Principal Component Analysis (PCA) as analyzed here, secondary to effective feature value evaluation. Prospective feature predictors are identified by using correlation testing with additional predictors. Those predictive earmarks together are some clinical specimens, some biochemical biomarkers, some imaging attributes, and many more. Once feature selection is in its best form, only the significant predictors will count in the prediction of the model. Applying ML to healthcare will have its share of challenges, despite the number of advantages that come with it. The first one worth highlighting is the quality of the data source and the issues with accuracy and availability of data. Medical data usually are incomplete, imbalanced, or noisy, which can lead to poor model performance. Moreover, concerns over data privacy, security, and ethics must be paid serious attention to in order to protect patient confidentiality [11]. Lack of interpretability and opacity are the most critical drawbacks in many ML models, specifically deep learning systems. Failure to explain the models' predictions may make some clinicians uncomfortable in putting a value on no explanation-based models for diagnostic or treatment-support tasks; hence, XAI must be incorporated. It is by means of rigorous validation, standardization, and integration with existing healthcare systems that the successful deployment of ML models in clinical practice is thus brought into fruition. The need at hand then is for the models to be carefully tested across acknowledged and validated populations and settings so that there is a guarantee of generalization and consistency. An atmosphere of co-operation among data scientists, clinicians, and healthcare stakeholders is a sine qua non for bridging such gap from tech lay down to practical implementation. To sum up, Machine learning has greatly improved the health care diagnosis through data-driven, precise, and efficient decision-making. Its capability to handle enormous datasets, recognize fresh disease patterns, and back up customized treatment presents itself as a major advantage for modern medicine [12]. However, those challenges in managing data quality, interpretability, and integration into

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

clinics are expected to be addressed to ensure that this asset is fairly recorded. In the years ahead where technology and research will surge ahead, Machine Learning is anticipated to become a major part of the post-transformation of healthcare diagnostics and patient outcome. Table 1 presents a concise overview of how Machine Learning techniques contribute to improving accuracy, efficiency, and decision-making in healthcare diagnosis.

Table 1: Role of Machine Learning In Healthcare Diagnosis

Category	**Approach/Model**	**Application in Healthcare Diagnosis**	**Advantages**	**Limitations**
Supervised Learning	Decision Tree, Random Forest, Support Vector Machine (SVM), Neural Networks	Used for disease classification and prediction based on labeled clinical datasets such as EHRs and lab reports	High accuracy, effective for structured data, widely adopted in diagnosis systems	Requires large labeled datasets, risk of overfitting
Unsupervised Learning	K-means Clustering, Hierarchical Clustering, Principal Component Analysis (PCA)	Identifies hidden patterns, patient stratification, and disease subtypes without labeled data	Useful for exploratory analysis, reduces dimensionality	Less precise for direct diagnosis, interpretation challenges
Reinforcement Learning	Q-learning, Deep Reinforcement Learning	Applied in treatment planning and adaptive clinical decision-making	Learns optimal strategies over time, dynamic adaptation	Complex implementation, limited real-world deployment
Deep Learning	Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN)	Widely used in medical imaging (MRI, CT scans, ultrasound) and sequential health data analysis	High performance in image and pattern recognition	Black-box nature, high computational cost
Feature Engineering	Feature Selection, Recursive Feature Elimination (RFE), PCA	Identifies key clinical and biochemical parameters (e.g., BMI, glucose,	Enhances model accuracy and reduces complexity	Risk of losing important information if not properly applied

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

		hormones) for improved prediction		
Predictive Analytics	Classification & Regression Models	Enables early disease detection and risk prediction for chronic diseases including PCOS, diabetes, and heart disease	Supports preventive healthcare and timely intervention	Dependent on data quality and completeness
Explainable AI (XAI)	SHAP, LIME	Provides interpretation of ML predictions, highlighting important features influencing diagnosis	Improves transparency and clinical trust	Additional computational overhead
Data Integration	Multi-source Data Fusion	Combines clinical, imaging, and lifestyle data for comprehensive diagnosis	Holistic analysis and improved decision-making	Data heterogeneity and integration complexity
Clinical Decision Support Systems	AI-integrated platforms	Assists clinicians in diagnosis and treatment recommendations	Reduces human error and improves efficiency	Requires validation and regulatory approval
Challenges in ML Adoption	Data Quality Issues, Class Imbalance, Privacy Concerns	Affects reliability and generalization of models in real-world healthcare systems	Highlights areas for improvement	Limits large-scale deployment

4. FEATURE SELECTION AND DATA PREPROCESSING TECHNIQUES FOR PCOS PREDICTION

Feature selection and data preprocessing are important steps in the making of accurate and reliable Machine Learning (ML) models for predicting Polycystic Ovary Syndrome (PCOS). Because PCOS is multi-faceted and complex, the datasets in prediction may contain elements such as clinical symptoms, hormonal profiles, metabolic indicators, lifestyle factors, and imaging data. These are usually high-dimensional, generally noisy, incomplete, and imbalanced data, which might extremely harm an ML algorithm's ability to do its work. Proper handling of these issues means that the possible ways to pre-process and choose a few features will be the basics in enhancing model accuracy, reducing computational complexity, and allowing for enhanced generalization and interpretability. Data pre-processing begins with data collection and integration from multiple sources such as hospital records, laboratory reports, and patient surveys. Since medical data is often prone to inconsistencies, the first critical step involves data cleaning [13]. This includes handling missing values, removing duplicate records, and correcting inconsistencies in data formats. Missing data is a common issue in healthcare datasets; techniques such as mean, median, or mode imputation can be used to address this. Advanced methods such as k-Nearest Neighbors (k-NN) imputation and multiple imputation may also be employed. One appropriate way to handle missing values is to make sure the dataset is still representative and lessen the possibility that our predictions are biased. Another essential pre-processing step is feature transformation and normalization. Medical features often exist on different scales, such as hormone levels, body mass index (BMI), and glucose levels, which may degrade the performance of ML models, especially those sensitive to feature scaling like Support Vector Machines and k-NN. Techniques like min-max normalization and z-score standardization are used to bring all features to a comparable scale to improve model convergence and performance [14]. Additionally, the variables categorical in nature, e.g., lifestyle habits or family history, must be turned into numerical form using encoding techniques like one-hot encoding or label encoding. Outliers are of high significance when it comes to the pre-processing task. Measurement errors, data entry mistakes, or rare medical conditions give rise to these outliers, distorting the learning process of the ML model. Popular methods for identifying and treating extra data points include using the interquartile range (IQR), z-score analysis, and clustering algorithms, depending on the problem. They may be treated differently-one might remove them, Transform them, or retain the outliers because they provide the specialist with the essential duties. Assuring good data quality at this stage is one of the cornerstones for building robust predictive models. The presence of imbalanced PCOS datasets become another roadblock, having a significant number of non-PCOS cases and only a few PCOS cases. This imbalance may lead to partiality in the model towards the majority class and, thus, a poor search for the minority class. To tackle this, data balancing is done. All upon any imbalanced metric considered, oversampling (e.g., SMOTE, Synthetic Minority Over-sampling Technique) and under sampling are implemented. They purposely create well-balanced datasets where both classes are equally presented to the model so that it learns best from patterns of either class. Cost-sensitive learning has been also employed to assign higher costs for misclassification of minority class instances. Feature selection is imperative, for it seeks to identify most relevant features that significantly contribute to the prediction of PCOS. With the elimination of irrelevant and redundant features, the model experiences an improved performance; without such elimination, the model may significantly overfit. Thus, this process also leads to an improvement in interpretability of the model. There are three broad categories for feature selection techniques: filter methods, wrapper methods, and embedded methods. In general, filter methods use statistical measures for the evaluation of features, including correlation coefficients, chi-square tests, mutual information, and ANOVA. These methods are statistically computed with a fairly low computation cost because they are independent of learning algorithms, and therefore they fare well for high-dimensional datasets [15]. In contrast, wrapper methods try to evaluate various subsets of features by training and testing a certain ML model. Recursive Feature Elimination (RFE), as well as forward and backward selection techniques, recursively select so that the model performance determines the optimal subset. Wrapper methods will produce more accurate predictions most oftentimes because they are usually computationally expensive and do not support large datasets. Embedded methods are one unique group where feature selection naturally arises within the model training. Lasso Regression, Ridge Regression, and tree-based models such as Random Forest inherently perform feature selection by assigning importance scores to features. Such methods are the best compromise between efficiency and accuracy. In the context of predicting the condition of PCOS, some of the most frequently chosen features are body mass index (BMI), insulin levels, glucose levels, follicle count, menstrual irregularity, age, hormonal ratios like LH to FSH, and lifestyle factors like physical activity and diet. In advanced feature engineering, new features could be derived given existing ones, like some of those that combine numerous hormonal indicators or ratios that are likely to give better predictive accuracy [16]. The importance-identification of each of these engineered features does not not only increase model performance but also furnish precious clinical knowledge by identifying the essential points affecting PCOS. Dimensionality reduction techniques like PCA and LDA are used to project data from their high input dimensionality into significantly low-dimensional spaces while preserving the key information. Reduction of noise, enhancement of computational efficiency, and simplification of complex data patterns are achieved by use of these techniques. Nevertheless, interpretability may be endangered as the transformed features represent combinations of original variables and do not have direct clinical interpretations. Another important step in preprocessing is the splitting and validation of data. Data splitting is normally performed into three separate sections – the training, validation, and test sets - for the evaluation of model performance and avoidance of any overfitting. Cross-validation techniques like k-fold cross-validation are usually employed to ensure that the model would yield consistent performance across different partitions of data. Proper validation is crucial for the assessment of reliability and generalizability of the predictive model.

5. MACHINE LEARNING MODELS FOR PCOS DETECTION

In recent years, machine learning (ML) models have gained attention for their potential to enhance the detection and diagnosis of Polycystic Ovary Syndrome (PCOS). Because PCOS possesses both complex and heterogeneous nature, the traditional diagnostic methods are sometimes burdened by lack of accuracy and sometimes by inconsistency. So, under such complexities, ML models can use extensive clinical, biochemical, and lifestyle data to recognize patterns and relationships that might not easily be detected by conventional methods. In turn, these models provide analytics-driven insights early in treatment either by observing progression or preventing early termination of a treatment when necessary, thereby reducing equivoques toward diagnosis and thus enabling personalized health solutions to PCOS. Amidst various ML techniques, utilizing supervised learning models for the detection of PCOS is the most popular approach. Decision Trees, Support Vector Machines (SVM), Random Forest, k-Nearest Neighbours (k-NN) and Artificial Neural Networks (ANN) are among the algorithms that depict good performance to classify. Decision Trees are very useful because they are simple, interpretable and provide a fine structure of decision rules based on input features. Random Forests, an ensemble learning method, take a step forward by combining many trees to enhance accuracy and decrease overfitting. It is especially useful when dealing with high-dimensional datasets and identifying key features affecting the PCOS situation. Support Vector Machines are another popular choice because they could manage nonlinear relationships, performing well on high-dimensional feature spaces [17]. With the help of kernel functions, SVM models showed impressive accuracy for the separation of PCOS and non-PCOS cases. Similarly, k-Nearest Neighbors classify new cases depending on the degree of their similarity to existing data points that this technique is very helpful for pattern recognition but still manageable for large datasets. Artificial Neural Networks, inspired by our very own brain, can create models for complex nonlinear relationships, and they have shown some promising results in PCOS prediction, particularly if the dataset is large. In addition to the conventional ML models, deep learning approaches to PCOS detection have been evaluated. Convolutional Neural Networks (CNNs) have shown particular promise regarding medical images such as ultrasound scans used to identify the morphology of a polycystic ovary. In images, such models automatically pick up what information is necessary while removing the need for manual feature engineering. Meanwhile, Recurrent Neural Networks are more congruous when dealing with sequential data, such as tracking changes in hormones over time. It should be noted that despite the high accuracy, deep learning models usually demand large datasets and significant computational resources.

Feature selection plays a critical role in improving the performance of ML models for PCOS detection. Common features that are used include body mass index (BMI), indices of insulin resistance, glucose levels, menstrual irregularity, follicle counts, and hormonal ratios, such as Luteinizing Hormone (LH) to Follicle Stimulating Hormone (FSH). Choosing the right features can enhance the accuracy of the models while keeping the computational wear and tear low. Ensemble techniques that use several machine learning models have been proven to enhance predictors by concatenating the experiences from several algorithms. Accordingly, hybrid ensembles that combine feature selection techniques with ensemble-learning techniques have shown to be particularly powerful in competition with standard single methods for predictions. Despite the positive outcomes, there are some considerable challenges which must be addressed in the use of ML models for PCOS detection [18]. An

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

issue of utmost concern is the availability of standard and quality data sets. The data is often incomplete, imbalanced with varying information, depending on the set of populations from which it comes. This could significantly affect the ML model's generalizability, limiting its use in everyday medical practice. Another critical issue is that many ML models are not interpretable, particularly with deep learning techniques, rendering lab interpretations difficult to comprehend and trust. In the future, more and bigger opportunities for data science will come as there is a rapid pace of growth in new technologies-faster systems and data sources, all starting to inform data scientists who need to stay ahead in this new world of healthcare.

Summarily, ML models provide a powerful, quick, and powerful way to detect PCOS by using data-driven insights combined with advanced computational techniques. From old-fashioned traditional classifiers to deep-learn AI structures, they showed tremendous potential in enhancing the existing diagnostic correctness regarding the early stages of the disease. However, aiming at solving the roadblock of data quality and interpretability, as well as closer integration with clinics, remains pivotal for really making this gulfab wizardry a great service for the average patient. To this end, there must be continuous research and innovation in order to bring forth such predictive systems for spotting PCOS that are certainly more reliable and useful in general clinical settings.

6. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) IN PCOS PREDICTION

Significant strides in XAI, while addressing the issue of PCOS prediction, present a significant milestone when pushing for ML-enhanced health care. Although current ML models obtain impressive accuracies in disease detection, their "black-box" nature has restricted the application of ML models within the medical context. As such, doctors require not only accurate predictions but also clear answers regarding how these predictions are made. This need is answered by XAI via enhanced transparency, interpretability, and accountability within ML-based systems, thus basking trust and informed decision-making. Interpreting XAI by all essential requirements allows us to understand the very working mechanisms of these systems. To twist the story, let us suppose an XAI approach could be structured to provide the commoner model results. Many elements might achieve these results; say putative or target hormonal levels, termico-indicators, lifestyle features, and clinical symptoms. It is important, for one thing, to characterize the varied components that bring on this data-attaching mechanism [19]. It seems reasonable that in this particular clinical context, the malfunctioning of principal elements could help to evaluate these test predictors. This possibly implies that powerful interacting factors ascertain the current progression and subsequent outcome of PCOS diagnostic criteria. Additionally, the utmost alertness into judgmental strategies brings the evidence performed by clinical judgment in order for each parameter to undergo diagnostic or treatment planning. White-box explanations can support model-specific or model-agnostic perspectives. The former ones are developed specifically for certain models, such as decision trees or linear regression, where interpretability makes part of the models themselves. For example, decision trees give clearcut rules easy to apprehend for practitioners. The model-agnostic approaches, conversely, can be used for absolutely any machine-learning model, including intricate deep-learning architectures. Popular methods for model agnosticism are Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) [20].

In simple terms, SHAP can be configured to interpret relationally one feature against another in different combinations. Hence, it is based on considering the participatory effect from all possible combinations of features to the targeted output layer. RTE shall further be explained in individual aspects regarding not only crossing with regard to algorithms, but they also shall lend insights into competing features crossing in semi-lucrative models. The coronation of XAI into PCOS prediction has a handful of benefits. Above all, it enhances model transparency, causing the health provider to have access to verify, and it assures that AI prediction aligns with medical information. Secondly, it makes the trust and acceptance of AI systems in clinical settings better. Thirdly, it builds a bridge to feature importance analysis, which can provide valuable insights into the causes and risk factors underlying PCOS. These insights can help structure tailored treatment and enhance patient outcome [21].

In practice XAI raises numerous challenges in healthcare in spite of its substantial potential benefits. The most critical issue relates to the trade-off between model complexity and interpretability. Very often, there is a decrease in the interpretability of highly accurate models, such as deep neural networks, whereas simplicity can be a risk for the performance of a model. XAI techniques are associated with the provocation of high computational cost or implementation time (keeping pace with interpretability) toward realizing fruitful AI for health. As the interpretability of models affects these methods' relations with our world, XAI thus needs to be taken with utmost haste, care, and scrutiny. Ensuring that opposite explanations are accurate, actionable, and applied in a sustained means will form the essential benchmark for the real-life implementation of XAI. XAI integration in clinical workflows is that other consideration that must be identified. Explanations must be loaded in a way that can easily be understood by healthcare professionals if and when there is a bottom line. Explaining interaction graphs and dials is another must [22]. In the meantime, there should be teamwork that involves the concerned data scientists and the clinical side in enabling better alignment of XAI outcomes with clinical requirements [23]. To sum up, the Explainable Artificial Intelligence (XAI) is essential in enhancing the transparency and trust in the application of AI in the prediction of PCOS cases. Through providing clear and interpretable reasoning for model decisions, it effectively covers ground between highly advanced computational numeric techniques and clinical applications. Through those means, trust is built and adoption is fostered, hence making broader understanding of PCOS developing factors attainable. The ongoing research in this particular direction stands to combine XAI with machine learning techniques offering the possibilities of creating robust and field-accepted diagnostic utilities that are more interpretable ones. Table 2 provides a comparative overview of existing PCOS studies based on methodologies, datasets, performance, and key contributions.

Table 2: Comparative Analysis of Existing PCOS Detection and Analysis Studies

Ref. No.	Method/Approach	Dataset/Features Used	Accuracy	Key Contribution
[1]	ML Models (RF, SVM)	Clinical & hormonal data	~94%	Improved PCOS diagnosis using ensemble ML techniques
[6]	AI-based phenotyping	Follicle size distribution	~92%	Automated ovarian phenotype classification
[22]	CNN + Transfer Learning	Ultrasound + biomarkers	~96%	Hybrid deep learning model for accurate diagnosis
[17]	Ultrasound + ML review	Imaging features	~91%	Comprehensive evaluation of imaging-based diagnosis
[7]	Ultrasound feature analysis	Reproductive-metabolic markers	~89%	Identified predictive ultrasound indicators
[2]	OCT imaging analysis	Ocular vascular features	~88%	Novel biomarker linkage with PCOS
[14]	Hormonal modeling	Endocrine parameters	~87%	Hormonal imbalance impact analysis
[9]	Network analysis	Psychological + clinical data	~86%	Mental health correlation in PCOS patients
[19]	Histopathological review	Tissue-level data	~85%	Disease progression understanding
[4]	Organoid-based modeling	Cellular models	~88%	Future diagnostic platform
[18]	Molecular analysis	ECM biomarkers	~84%	Role of matrix remodeling
[20]	Pharmacological modeling	Bile acid regulation	~83%	Metabolic pathway insights
[24]	AI + phytochemical study	Ovarian morphology	~89%	Integrated therapeutic-diagnostic approach
[21]	AI-based nutrition model	Lifestyle & diet data	~87%	Nutritional impact on PCOS prediction

7. CONCLUSION AND FUTURE DIRECTIONS

In this review, the complexity and clinical significance of Polycystic Ovary Syndrome (PCOS) have been comprehensively analyzed, with a particular focus on the role of Machine Learning (ML) and Explainable Artificial Intelligence (XAI) in improving its diagnosis and prediction. PCOS continues to pose substantial challenges due to its heterogeneous nature, diverse symptomatology, and lack of standardized diagnostic procedures. Traditional diagnostic approaches, while widely used, often fail to provide timely and accurate detection, thereby increasing the risk of long-term health complications.

This review has emphasized the effectiveness of ML-based approaches in addressing these limitations by enabling data-driven and accurate prediction of PCOS. Various models, including Decision Trees, Random Forest, Support Vector Machines, and Neural Networks, have demonstrated strong performance in identifying patterns within complex medical datasets. The importance of data pre-processing and feature selection has also been highlighted, as these steps significantly influence model accuracy, efficiency, and generalizability. By focusing on relevant clinical and biochemical features, ML models can provide reliable predictions and support early intervention strategies. Furthermore, this review has explored the integration of XAI techniques to enhance the transparency and interpretability of ML models. Methods such as SHAP and LIME play a crucial role in explaining model decisions by identifying the contribution of individual features. Future research should focus on developing hybrid and ensemble models, integrating multimodal data sources, and leveraging advanced techniques such as deep learning and federated learning to enhance predictive capabilities while maintaining data privacy.

References

[1] Abed, Baraa Mohammed, et al. "Enhancing Polycystic Ovary Syndrome (PCOS) Diagnosis Using Machine Learning Techniques." INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES (2026): 2130-2139.

[2] Icoz, Mehmet, Melisa Golgelioglu, and Eyup Erkan. "Choroidal vascularization in polycystic ovary syndrome: an optical coherence tomography study." International Ophthalmology 46.1 (2026): 79.

[3] kumar Patel, Pankti Bhavik, et al. "A Comprehensive Review of Polycystic Ovary Syndrome (PCOS): Pathophysiology, Diagnosis and Management." Asian Journal of Pharmaceutical Research and Development 14.2 (2026): 32-41.

[4] Singh, Ishanka, Anuja Pant, and Pawan Kumar Maurya. "Organoids in polycystic ovary syndrome: Emerging platforms and future directions." Reproductive Biology 26.3 (2026): 101217.

[5] Ni, Xiaorong, et al. "Reactive Oxygen Species-responsive release of tanshinone IIA for polycystic ovary syndrome treatment." (2026).

[6] Agirsoy, Mehtap, and Matthew A. Oehlschlaeger. "AI-Based Ovarian Phenotyping Using Follicle Size Distribution Patterns for PCOS Assessment." (2026).

[7] Pourkhani, Zahra, et al. "The value of ultrasonographic features in diagnosis and prediction of reproductive-metabolic dysfunctions in polycystic ovary syndrome: a narrative review." Middle East Fertility Society Journal 31.1 (2026): 19.

[8] Bal Taştan, Tuğba, et al. "Astaxanthin and omega-3 fatty acids enhance endometrial receptivity in a rat model of polycystic ovary syndrome." Naunyn-Schmiedeberg's Archives of Pharmacology 399.6 (2026): 8311-8326.
[9] Li, Shaojing, et al. "Mapping the network structure of anxiety, depression, and sleep symptoms in patients with polycystic ovary syndrome." Frontiers in Psychiatry 17 (2026): 1738355.
[10] Momin, Ruksar A., et al. "EFFECTIVENESS OF STRUCTURED TEACHING PROGRAMME REGARDING POLYCYSTIC OVARIAN SYNDROME AMONG ADOLESCENT GIRLS IN A RURAL AREA OF MAHARASHTRA." International Journal of Medicine & Public Health 16.1 (2026): 2693.
[11] Chaurasia, Shweta, and Moumita Das. "Perception of Polycystic Ovary Syndrome (PCOS) Among Women: A Comparative Analysis Based on Developed and Developing Countries." The Palgrave Handbook of Global Social Problems. Cham: Springer Nature Switzerland, 2026. 1-17.
[12] AL-belazi, Balsam J., Lujain M. Aldiab, and Fathi M. Sherif. "Inositol has a beneficial effect in the management of polycystic ovary syndrome: Special data for Libyan patients." Mediterranean Journal of Medical Research 3.1 (2026): 14-26.
[13] Shah, Fatima Iftikhar, et al. "From Seed to Therapy: Sesamin and Phytosterols of Sesamum indicum in Polycystic Ovary Syndrome and Ovarian Cancer." Reproductive Sciences (2026): 1-25.
[14] Pandey, Sarita, et al. "Progressive impact of polycystic ovary syndrome on hormonal alterations, bone microarchitecture, and mechanical properties." Journal of Endocrinology 268.1 (2026).
[15] Abou-El-Naga, Amoura M., et al. "Date palm pollen extract improves fertility and pregnancy outcomes in letrozole-induced polycystic ovary syndrome rats." Journal of Molecular Histology 57.2 (2026): 75.
[16] Suster, David, et al. "Normal Structure and Function, Inflammatory and Other Miscellaneous Lesions of the Ovary." Gynecologic and Obstetric Pathology. Singapore: Springer Nature Singapore, 2025. 1105-1130.
[17] Di Michele, Stefano, et al. "Ultrasound assessment in polycystic ovary syndrome diagnosis: from origins to future perspectives—a comprehensive review." Biomedicines 13.2 (2025): 453.
[18] Nikanfar, Saba, and Christiani A. Amorim. "The potential role of matrix metalloproteinase in polycystic ovary syndrome: implications for extracellular matrix remodeling." Biofactors 51.3 (2025): e70026.
[19] Javed, Mehjbeen, et al. "Chemically-induced histopathological progression in the polycystic ovary syndrome (PCOS)-A review." Toxicological Research 41.5 (2025): 437-453.
[20] Tong, Tian-Tian, et al. "Pharmacological effects of bile acids on polycystic ovary syndrome via the regulation of chemerin." Chinese Medicine 20.1 (2025): 45.

[21] Yingngam, Bancha, Chakkrapong Chaiburi, and Abhiruj Navabhatra. "Introduction to Polycystic Ovary Syndrome (PCOS) and Nutritional Challenges." AI-Based Nutritional Intervention in Polycystic Ovary Syndrome (PCOS). Singapore: Springer Nature Singapore, 2025. 1-56.

[22] Sundari, M. Shanmuga, et al. "Transfer learning-enhanced CNN model for integrative ultrasound and biomarker-based diagnosis of polycystic ovarian disease." Scientific reports 15.1 (2025): 34519.

[23] Atta, Rasha, et al. "Pluchea dioscoridis extract: A novel therapeutic approach for polycystic ovary syndrome targeting ovarian morphology, dopamine pathways, and neurobehavior in relation to its phytocomponents." Plos one 20.6 (2025): e0325439.

[24] Singh, Dilpreet. "Nanotechnology-based Diagnostic Approaches for Early Detection and Monitoring of Polycystic Ovary Syndrome (PCOS)." Current Analytical Chemistry 21.4 (2025): 263-275.