RT&A, No 4 (89)
Volume 20, December 2025

PREDICTIVE MODELING OF POLYCYSTIC OVARY SYNDROME USING MACHINE LEARNING ALGORITHMS

Gneya Pandya[1], Dhara Solanki[2], Dr. Jigna Jadav[3], Dr. Kajal Patel[4]

•

[1,2]Student, Computer Engineering, V.G.E.C., Chandkheda,Gujarat, India.
[3]Assistant Professor, Computer Engineering, V.G.E.C., Chandkheda, Gujarat, India.
[4]Associate Professor, Computer Engineering, V.G.E.C., Chandkheda, Gujarat, India.
gneyapandya1234@gmail.com[1], dhara732002@gmail.com[2],jigna.j.jadav@gmail.com[3], kajalpatel@vgecg.ac.in[4]

Abstract

PCOS is a prevalent health condition affecting women globally, and if not diagnosed early, it can lead to severe complications like type two diabetes and gestational diabetes. Our research is focused on enhancing the accuracy and reliability of PCOS diagnosis using machine learning techniques. We utilized a public dataset and applied a range of machine learning models like Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) ensemble algorithms. By implementing feature selection methods, we aim to identify the most critical factors contributing to the diagnosis, ensuring that the best-performing model is selected for reliable results. Notably, our research achieved a very good accuracy by using the Random Forest algorithm, highlighting its potential to provide a more precise and trustworthy diagnosis of PCOS. This work represents a significant step toward improving healthcare outcomes for women with PCOS.

Keywords: *Polycystic Ovary Syndrome (PCOS), Machine Learning, Feature Selection, Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Model Interpretability, Ensemble Learning, Exploratory Data Analysis (EDA).*

I. Introduction

PCOS is a prevalent hormonal disorder affecting fertile women [1]. PCOS often begins during teenage, but its symptoms can vary over time. Hormonal abnormalities, irregular menstrual cycles, increased testosterone levels, and ovarian cyst development are among the possible outcomes. Unusual menstrual cycles, typically caused by insufficient ovulation, can pose difficulties in conceiving. PCOS is a chronic disorder that is not curable, but specific symptoms can be treated with medication, lifestyle changes, and fertility therapies. It is one of the leading causes of infertility. Even though the precise origin of PCOS is yet unknown, women who have type 2 diabetes or a family history are more susceptible. One of the most prevalent hormonal disorders affecting women who can conceive, PCOS is a significant public health concern. An estimated 8–13% of women who can conceive are affected by the illness, and up to 70% of instances go misdiagnosed [1].

Diabetes, heart disease, high blood pressure, endometrial thickness, sleep apnoea, depression, anxiety, eating disorders, and endometrial cancer are among the conditions that women with PCOS are likely to experience [2]. Environmental variables may also influence the development of PCOS in addition to hereditary ones. You can lessen long-term consequences in addition to losing

DOI: https://doi.org/10.24412/1932-2321-2025-489-172-179

weight, getting treatment, and getting diagnosed early [2].

AI can now identify and cure complicated illnesses [3]. AI techniques create a robust PCOS diagnostic framework by extracting information from massive data sets using state-of-the-art algorithms and potent machine learning [4]. While some predictive models only include women who seek reproductive treatment, they can be quite helpful in supporting the early identification of PCOS. One model predicted the development of PCOS in Chinese women based on BMI, menstrual cycle duration, and serum levels of androstenedione and anti-Müllerian hormone (AMH). Another model predicted a diagnosis of PCOS or other ovulatory dysfunction illnesses based just on AMH and BMI. Predictive models for specific outcomes, like pregnancy outcomes and insulin resistance, have been developed by different research involving women with PCOS [5]. One popular solution to reduce the issue of too many unnecessary features is feature selection. A well-chosen feature set can decrease the complexity of the learnt outcomes while improving predicting accuracy and learning efficiency [6]. It entails locating and picking the most pertinent variables from a dataset to enhance a machine learning model's functionality and readability. Feature selection reduces complexity, minimizes overfitting, and improves the model's capacity to generalize from the data by concentrating on the most significant features.

EDA helps to identify which variables are redundant and which contribute to significant variation in the lower dimensional space by identifying and characterizing recurrent patterns and significant correlation structures [7]. EDA provides a foundational understanding of the data, helping to inform decisions about which features to select and how to approach the modelling process. An important subfield of artificial intelligence, machine learning, was discovered in the 1950s [8]. Machine learning algorithms utilize EDA's selected features and insights to build predictive models. These algorithms apply various statistical and computational techniques to learn from the data, making predictions or classifications based on the input features. The effectiveness of these algorithms often hinges on the quality of feature selection and the depth of insights obtained through EDA. Various Machine Learning algorithms are available, which include SVM, LR, DT, etc.

In this study, we developed and evaluated multiple machine learning models—including Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF)—to enhance the accuracy of Polycystic Ovary Syndrome (PCOS) diagnosis. A comprehensive comparative analysis was conducted to assess the performance and diagnostic effectiveness of these models. Additionally, feature selection techniques were employed to identify and isolate the most relevant features significantly contributing to the prediction of PCOS, thereby improving model interpretability and diagnostic reliability.

The structure of this document is as follows. Section II provides an overview of the field's current state of the art. Section III provides an account of the research conducted in this study. Section IV presents the findings from the analysis of the dataset. The conclusions can be found in section V.

II. Related Work

In their work, Authors in [9] tackled the problem of PCOS detection, a prevalent endocrine condition that affects around 15% of women worldwide. In their work, PCONet—a CNN created to identify PCOS in ovarian ultrasound pictures—was presented. Additionally, they used transfer learning to refine InceptionV3, a 45-layer pre-trained CNN, for categorization. According to the findings, PCONet outperformed the refined InceptionV3, which achieved 96.56% accuracy, with a gain of 98.12%. This study shows how transfer learning and sophisticated CNN models can improve the accuracy of PCOS identification.

Authors in [10] looked into the predictive potential of machine learning algorithms for PCOS, a common endocrine condition that affects women of reproductive age. Their study used a dataset with variables like Age, body mass index (BMI), hormone levels, monthly irregularities, and

lifestyle factors to assess different machine learning algorithms, such as decision trees, random forests, and linear regression. The project aims to create a reliable predictive model that will help medical professionals identify PCOS early on. The results highlight how machine learning may help women's health by facilitating early detection and individualized care, making a substantial contribution to the endocrine problem predictive modeling field.

Using T2-weighted Magnetic Resonance Imaging (MRI) radiomics analysis [11] investigated how well T2-weighted MRI could distinguish between classical and non-classical forms of PCOS. The dataset for the study was split into training and test sets, and it comprised 202 ovaries from 101 PCOS patients. Using machine learning algorithms, the investigation showed that the most successful methods were Random Forest and Gradient Boosting Classifier, which produced accuracies of 73% and 70%, respectively. These algorithms' blend model yielded an AUC of 0.70 and an accuracy of 73%. The results emphasize the potential of machine learning and radiomics in PCOS diagnosis by demonstrating the use of radiomic characteristics from T2-weighted MRI in differentiating between PCOS phenotypes.

These advancements underline the transformative impact of artificial intelligence in medical diagnostics, particularly for complex endocrine disorders like PCOS. By integrating clinical data, imaging techniques, and robust machine learning models, researchers are moving closer to developing automated, non-invasive, and highly accurate diagnostic tools. Such innovations not only aid in early detection but also enable personalized treatment planning, thereby improving patient outcomes. Future research may further explore hybrid approaches combining clinical features with radiomic and genomic data to build even more precise and comprehensive PCOS diagnostic systems.

III. Proposed Approach

Figure 1 illustrates a detailed workflow designed for predicting Polycystic Ovary Syndrome (PCOS) using a machine learning-based approach, specifically the Random Forest Classifier. The process begins with the collection of a well-curated PCOS dataset, which forms the foundational input for model training. This dataset then undergoes a crucial stage of feature engineering, wherein relevant variables are systematically identified, transformed, and optimized to improve the model's learning efficiency. This step is instrumental in ensuring that only the most impactful features are retained, thereby enhancing the predictive accuracy of the model while also minimizing computational overhead.

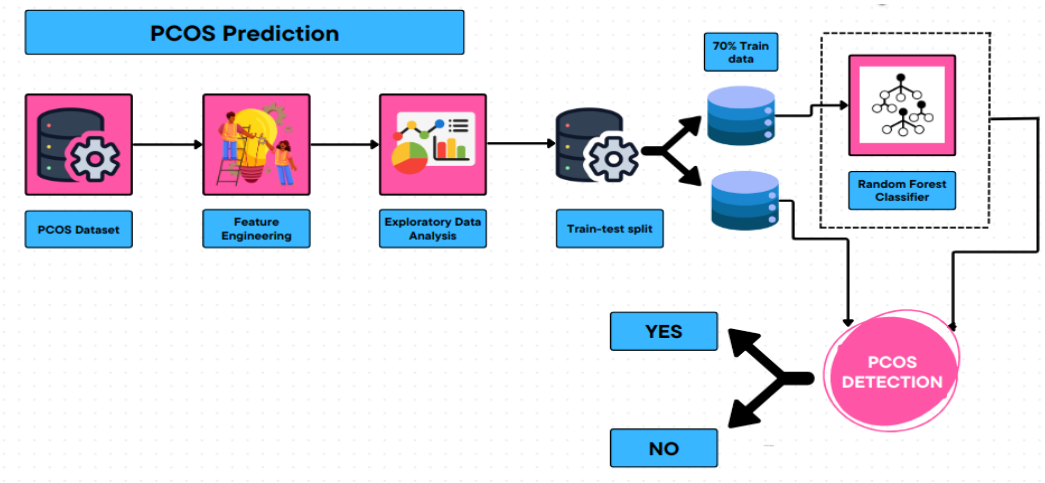

Figure 1. *Workflow of the PCOS Prediction Model using Random Forest Classifier*

Afterwards, an EDA is carried out to obtain a profound understanding of the dataset. Underlying trends, correlations, and other anomalies that can affect the model's performance are found using EDA. It also offers guidance for additional feature improvement and improves comprehension of the distribution and properties of the data.After that, the dataset is divided into subgroups for training and testing, with 70% of the data going into the former. This segmentation is essential for assessing the model's capacity generalization. The RF, an ensemble learning technique renowned for its resilience and efficiency in managing intricate, non-linear relationships within the data, is constructed using the training subset. During training, the classifier builds several decision trees and aggregates the results to produce a final prediction.The trained model is then applied to new data to predict the presence of PCOS. Based on the input features, the mode categorizes the results as either "PCOS Detected" or "PCOS Not Detected." This workflow highlights the systematic approach to PCOS detection and the potential of machine ML techniques in enhancing diagnostic accuracy, contributing early and more precise medical interventions for individuals at risk of PCOS.

I. Methodology

PCOS stands as one of the most prevalent hormonal disorders affecting women globally, with a complex set of symptoms and potential health implications. PCOS often manifests in symptoms like irregular period cycles, hormonal imbalances, and multiple cysts on the ovaries. It is closely associated with various health complications, including infertility, diabetes, cardiovascular diseases, and emotional well-being challenges Recognizing its widespread impact, this research project addresses PCOS from multiple dimensions, amalgamating machine learning techniques.

II. Data Collection

Our research study utilized a comprehensive dataset from ten distinct healthcare facilities across Kerala, India. The PCOS dataset [12], comprises clinical and physical parameters from 541 patients and encompasses 44 features.

III. Data Preprocessing

To prepare the combined dataset for analysis, we performed thorough data preprocessing. This involved addressing missing values by imputing them with zeros and eliminating extra information such as the 'Sl. No' and 'Patient File No.' columns from the PCOS dataset.

IV. Feature Selection

After merging the datasets, we conducted feature selection to identify the most relevant variables for our analysis. This step involved evaluating the importance of each feature in predicting the presence of PCOS and determining which variables would be included in our predictive models.

Following features were identified as influential predictors of PCOS in our models: ['PCOS (Y/N)', ' age (yrs)', 'Pregnant(Y/N)', 'No of aborptions', 'Bloated', 'facial hair', 'chest hair', 'difficult to loose weight', 'mood swings', 'anxiety/depression/stress', 'Irregular_sleep', 'Weight gain(Y/N)', 'hair growth(Y/N)', 'Skin darkening (Y/N)', 'Hair loss(Y/N)', 'Pimples(Y/N)', 'Fast food (Y/N)', 'Reg Exercise(Y/N)', 'Weight (Kg)', 'Height(Cm) ', 'BMI', 'Blood Group', 'Pulse rate(bpm) ', 'Cycle(months)', 'Cycle length(days)', 'Marriage Status (Yrs)', 'Hip(inch)', 'Waist(inch)', 'Waist/Hip Ratio']

V. EDA

We performed EDA on the combined dataset to acquire insights into the distribution and relationships among different features. This involves visualizing the data through histograms, scatter plots, correlation matrices, and other statistical summaries to identify patterns and trends.

Figure 2 presents a correlation heatmap showing the interrelationships among various clinical and physiological features related to PCOS. The results indicate that PCOS is positively correlated with skin darkening, weight gain, hair growth, and pimples, suggesting these symptoms are more prevalent in affected individuals. Additionally, strong correlations are observed between BMI, waist size, and weight, indicating redundancy among obesity-related features. Conversely, age and cycle length show weak negative correlations with PCOS, implying a slight decrease in PCOS prevalence with increasing age and shorter cycles.

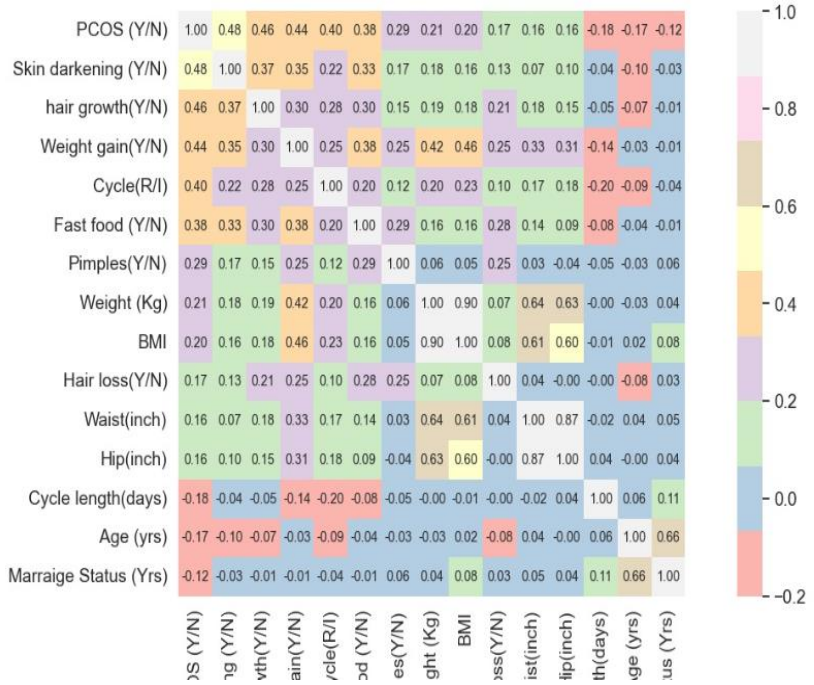

Figure 2. *Heatmap describing the correlation between features*

Figure 3 illustrates a scatter plot comparing cycle length and age for individuals with and without PCOS. It is evident that women with PCOS experience greater variation in cycle length, often showing irregular or prolonged cycles outside the typical 4 to 6-day range. The regression lines for both groups are nearly flat, indicating that age has minimal influence on cycle length. However, the increased dispersion among PCOS cases reinforces the relevance of cycle irregularity as a significant diagnostic indicator. These findings support the significance of cycle irregularity and hormonal symptoms as important indicators in PCOS detection.

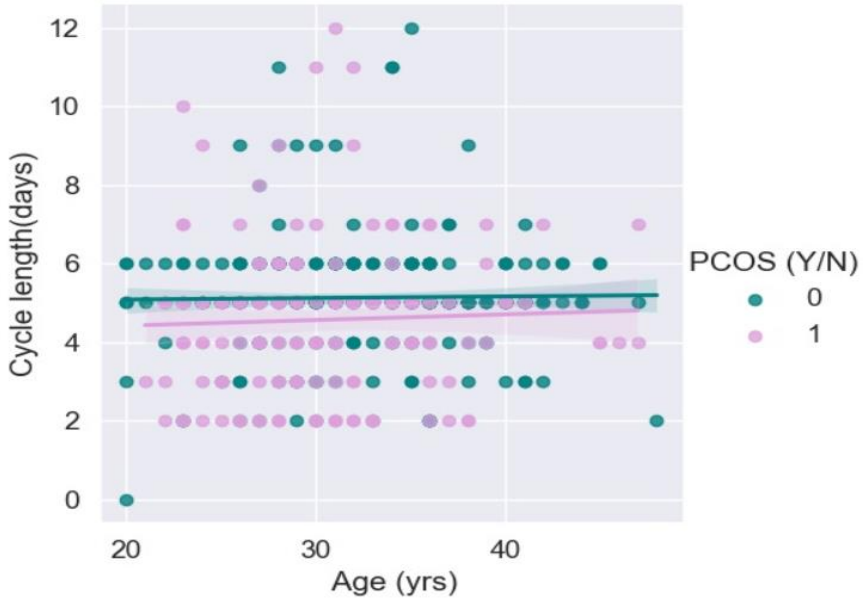

Figure 3. *Scatter Plot of Menstrual Cycle Length vs. Age about PCOS Diagnosis*

IV. Experimental Results and Discussion

I. Linear Regression

LR is a simple and effective method for binary classification tasks like PCOS detection. Its ability to provide probabilities and interpretability makes it an ideal choice for PCOS detection [13].

$$\text{logit}(Y) = a + B1X1 + B2X2 \quad (1)$$

LR achieved a respectable accuracy of 79% in predicting PCOS likelihood. This model's performance signifies its utility in offering users actionable insights about their PCOS risk based on provided data.

II. Decision Tree

DT were employed for their ability to create intuitive, interpretable models. Despite a slightly lower accuracy of 67%, decision trees are valuable resources for understanding how PCOS is influenced. The transparency of the decision tree structure aids users in understanding the key predictors of PCOS. The mathematical formula of decision trees uses Entropy E(S) .

$$E(S) = -Xn\ i=1\ pi \log2\ (pi) \quad (2)$$

The final formula is to calculate information gain [14]:

$$IG = E(S) - X\ v\in\ Values(A)\ |Sv|\ |S|\ E(Sv) \quad (3)$$

III. Support Vector Machine

SVM were selected for their ability to handle complex relationships in high-dimensional spaces. With a moderate accuracy of 59%, SVM provides an additional perspective on PCOS likelihood. The focus is on capturing intricate patterns within the data [15].

The mathematical formula of SVM is:

$$f(x) = \text{sgn}[\ SV\ s\ \alpha iyi(xi)\ Tx + b] \quad (4)$$

IV. Random Forest

An RF is created by combining different tree predictors so that every tree in the forest depends on the values of a random vector that is randomly sampled and has the same distribution for every tree [16]. Random Forests, as an ensemble learning method, were selected for their ability to improve predictive accuracy. With an impressive accuracy of 83%, the random forest model excels in capturing complex relationships within the PCOS dataset, providing users with highly accurate risk assessments.

V. Comparision of Models

Table 1. *Comparison of models based on accuracy, strength, and limitations*

Model	Accuracy	Strengths	Limitations
LR	79%	Simplicity and interpretability	Assumes linear relationships between features
DT	67%	Intuitive and easy to interpret	Prone to overfitting, especially with deep tree
SVM	59%	Effective in high-dimensional spaces	Highly computational, mainly when working with big datasets
RF	83%	combining multiple decision trees	Less interpretable compared to individual decision trees

V. Conclusion and Future Work

As a result of this research, machine learning techniques have contributed significantly to improving the diagnosis of PCOS. We have shown how such methods can increase diagnostic precision by utilizing and comparing several models, including LR, SVM, DT, and RF. Utilization of feature selection techniques has further refined our models, with the Random Forest algorithm achieving the highest accuracy of 83%.This outcome highlights the effectiveness of advanced machine learning approaches in providing more reliable and accurate PCOS diagnosis. Our findings contribute to the ongoing efforts to manage PCOS better and offer a foundation for future research to improve diagnostic methods and ultimately enhance patient outcomes.

In future research, integrating Internet of Things devices like Galvanic Skin Response sensors could provide valuable insight into the emotional and physiological states of women with PCOS. Women suffering from PCOS often experience heightened anxiety and depression. By incorporating GSR data with PCOS diagnostic models, we could develop more personalized treatment strategies to address physiological and psychological aspects. Additionally, sentiment analysis for mood tracking could enhance PCOS management by analyzing written communications to identify mood patterns and triggers related to symptoms. This approach could improve our

understanding of the emotional factors influencing PCOS, leading to more tailored and practical support and treatment options.

References

[1]G. Jiskoot, Y Louwers, A. Van der Mijle, V. Maas.(2025). Medication Use in Pregnant Women with Polycystic Ovary Syndrome (PCOS): A Nationwide Cohort Study. *Human Reproduction,*40,Deaf 097-948.

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

[3] M. M. Rahman.(2022). A web-based heart disease prediction system using machine learning algorithms, *Network Biology*, vol. 12, no. 2, p. 64

[4] A. Dar, M. Maqbool, Z. Qadrie, I. Ara, and A. Qadir.(2024).Unravelling PCOS: Exploring its causes and diagnostic challenges, *Open Health*, vol. 5, pp. 1–8

[5] Z. Zad, V. S. Jiang, A. T. Wolf, T. Wang, J. J. Cheng, I. C. Paschalidis, and S. Mahalingaiah.(2024). Predicting polycystic ovary syndrome with ML algorithms from electronic health records,*Frontiers in Endocrinology*, vol. 15, p. 1298628

[6] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan.(2022).A review of feature selection methods for machine learning-based disease risk prediction, *Frontiers in Bioinformatics*, vol. 2, p. 927312

[7] V. Da Poian, B. Theiling, L. Clough, B. McKinney, J. Major, J. Chen, and S. Hörst.(2023).Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry,*Frontiers in Astronomy and Space Sciences*, vol. 10, p. 1134141

[8] Ö. Çelik.(2018).A research on machine learning methods and its applications,*Journal of Educational Technology and Online Learning*, 1(3), 25-40.

[9] A. K. M. S. Hosain, M. H. Kabir Mehedi, and I. E. Kabir.(2022). Pconet: A convolutional neural network architecture to detect polycystic ovary syndrome (PCOS) from ovarian ultrasound images,in *Proc. Int. Conf. Engineering and Emerging Technologies (ICEET)*, IEEE.

[10] M. Priyadharshini, A. Srimathi, C. Sanjay, and K. Ramprakash.(2024).PCOS disease prediction using machine learning algorithms, *International Research Journal on Advanced Engineering Hub* , vol. 2, pp. 651–655.

[11] G. Rona .(2024).Machine learning-based analysis of MRI radiomics in the discrimination of classical and non-classical polycystic ovary syndrome, *cukurova medical journal,*vol. 49, no. 1, pp. 89–96.

[12] P. Kottarathil.(2024). *Polycystic ovary syndrome (PCOS)*, [Online]. Available: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos, ver. 1.0.

[13] J. Peng, K. Lee, and G. Ingersoll.(2002). An introduction to logistic regression: Analysis and reporting, *The Journal of Educational Research*, vol. 96, pp. 3–14.

[14] D. Mienye, N. Jere.(2024).A survey of decision trees: Concepts, algorithms, and applications, *IEEE Access*, vol. 12, pp. 86716–86727.

[15] H. Wang, J. Xiong, Z. Yao, M. Lin, J Ren.(2017). Research survey on support vector machine. InProceedings of the 10th EAI international conference on mobile multimedia communications 2017 Dec 8 (pp. 95-103).

[16] L. Breiman (2001), Random forests, *Machine Learning*, vol. 45, pp. 5–32.