A Machine Learning Approach for Detecting Polycystic Ovary Syndrome using XGBoost

Prathamesh Bhosale[1], Shashikant V. Athawale[2], Balkrishna Bile[3], Samarth Biraje[4], Niharika Dangat[5]

[1]AISSMS College of Engineering, Savitribai Phule Pune University, Pune, India

[2]Associate Professor, Department of Computer Engineering, AISSMS College of Engineering, Savitribai Phule Pune University, Pune, India

[3] AISSMS College of Engineering, Savitribai Phule Pune University, Pune, India

[4] AISSMS College of Engineering, Savitribai Phule Pune University, Pune, India

[5] AISSMS College of Engineering, Savitribai Phule Pune University, Pune, India

ABSTRACT

Abstract - Polycystic Ovary Syndrome(PCOS) is an endocrine disorder which affects women and girls of reproductive age globally. It causes multiples effects such skin pigmentation, hair loss, hair growth, etc, major being infertility and onset of mentals disorders such as anxiety and depression. Early prediction of PCOS can reduce complexity in treatment, therefore there is a need for proper PCOS prediction system because of its widespread occurrence and severity when remained unchecked. Present methods for the accurately diagnosis of PCOS are time consuming, costly and sometimes inconclusive. This research paper presents our work that would incorporate data science technique and machine learning into the PCOS diagnosis process to make it more accessible to the people and an help to doctors. Here in this study we use an ensemble machine learning classification Extreme Gradient Boosting (XGBoost) Classification technique for PCOS identification with patient's symptom data. The dataset is trained and tested with 70:30 ratio using utilizing different features. As outcome the proposed ensemble technique significantly increases the accuracy when compared to other ML techniques. Our research helps medical community and provide cure to women through early prediction of PCOS.

keywords : polycystic ovary syndrome, PCOS, XGBoost, machine learning, women's health, endocrine disease, data science, predictive modelling, medical diagnosis

1. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is one of the most common endocrine disorders in women. It is a hormonal condition that affects women of reproductive age. PCOS effects about 6% to 13% of women, out of which up to 70% of them remain undiagnosed.[1] [2]

Clinical signs of PCOS, including irregular menstruation, excessive hair growth, and cystic ovaries, were first noted by Stein and Leventhal in 1935[3]. Nonetheless, ovarian abnormalities have been observed since the 18th and 19th centuries[4]. Over time, studies have shown that PCOS is a complex disorder with genetic, metabolic, and hormonal factors, including insulin resistance[5][6].

Ovulatory dysfunction, hyperandrogenism, and polycystic ovarian morphology are the three features that must be present for the widely used 2003 Rotterdam criteria to be met[7]. These criteria evolved as a result of diagnostic challenges brought on by symptom variability. More recently, developments in artificial intelligence and molecular biology have improved methods for diagnosis and individualized treatment[8][9]. The intricacy and continuously evolving understanding of PCOS as an endocrinological condition are highlighted by the last century of research.

Associated health problems:

PCOS doesn't only affect the ovaries, it also causes other health problems, such as:

Infertility: This can happen due to irregular ovulation.

Obesity: This can be a symptom to many of the women suffering from PCOS.

Metabolic syndrome: It is a group of conditions like high blood pressure, high blood sugar and high cholesterol.

Type 2 diabetes: This can happen due to resistance to insulin over a long period of time.

Heart disease: This can be caused due to high blood pressure & high cholesterol in the body.

Depression & anxiety: It can also affect the mental health because of the imbalance in the hormonal health as well as the body image concerns.

Endometrial cancer: This can happen due to prolonged unopposed estrogen exposure (irregular shedding of uterine lining).

1.1 Research Objectives:

The objective is to study a machine learning approach that helps to detect the symptoms of PCOS at an early stage, which can help the females & also to spread awareness about this disorder. The ML approach would first gather the patient's basic health data such as age, height, weight, BMI, blood group, menstrual cycle, etc and then the PCOS status is shown.

1.2 Problem statement:

PCOS remains underdiagnosed and often misdiagnosed, causing delayed treatment and severe health implications. The current diagnosis process is highly dependent on manual interpretation and access to specialized medical infrastructure, limiting early detection in rural and underserved areas. There is a need for a cost-effective, intelligent system capable of analyzing patient health parameters such as hormonal levels, menstrual patterns, BMI, and lifestyle factors to provide reliable predictions, supporting healthcare professionals in early intervention.

1.3 Significance & motivation:

Unawareness about PCOS and the delay to detect the disorder in women can increase the risk of infertility, type 2 diabetes, heart disease, etc. Increasing awareness among people in the society can lead to early detection of PCOS. This can help to decrease the number of cases of PCOS that often go undiagnosed that is currently up to 70%[2].

2. LITERATURE REVIEW

Year	Author	Objective	Contribution	Data	Methodology	Result
2025	Kamal Upreti et al.	Enhance PCOS clinical diagnosis accuracy using AI	Showed hybrid AI models (SWISS-AdaBoost, ensemble) improve early detection	Various datasets including hormonal profiles, imaging	Ensemble learning, AdaBoost, SVM, RF, CatBoost, deep learning	Achieved accuracy up to 98%, AUC 99%, early diagnosis, reduced diagnostic delays

2025	Mehtap Agrisoy & Mathew A. Oehlschlaeger	To develop a non-invasive, accurate and cost-effective PCOS diagnostic model using clinical and ultrasound features through machine learning techniques.	Proposed an XGBoost-based ML model integrating ultrasound and clinical data for PCOS diagnosis; demonstrated that combining these features enhances diagnostic accuracy and efficiency.	Dataset from 541 women (185 PCOS, 356 non-PCOS) collected from 10 hospitals in India; includes clinical, biochemical, and ultrasound features (41 total).	Compared ANN, SVM, LR, KNN, XGBoost algorithms; used feature selection (Chi-Square SelectKBest, XGBoost feature importance, SHAP analysis); applied SMOTE, 10-fold cross-validation, and hyperparameter tuning.	Best model: XGBoost; with clinical + ultrasound features + AMH: AUC = 0.9947, Precision = 0.9553, F1 = 0.9553, Accuracy = 0.9553.

Upreti et al. (2025) aimed to enhance the clinical diagnostic accuracy of PCOS by integrating advanced artificial intelligence (AI) methodologies. Their study proposed hybrid ensemble frameworks such as SWISS-AdaBoost, combining multiple classifiers (AdaBoost, SVM, RF, CatBoost, and deep learning architectures) to improve predictive reliability. Drawing data from diverse clinical sources—including hormonal profiles and imaging datasets—the hybrid model achieved an accuracy of up to 98% and an AUC of 0.99. The results demonstrated that ensemble-based AI systems could not only improve early PCOS detection but also reduce diagnostic delays, highlighting their potential integration into routine clinical workflows for women's health diagnostics.

Agirsoy et al. (2025) focused on developing non-invasive, machine learning models for PCOS diagnosis using multimodal data. The dataset consisted of 541 patients (PCOS and non-PCOS) collected across ten hospitals in India, encompassing clinical, biochemical, and ultrasound attributes. Multiple algorithms—including Artificial Neural Networks (ANN), SVM, Logistic Regression, K-Nearest Neighbors (KNN), and XGBoost—were tested with feature selection techniques such as SHAP and SelectKBest to optimize model interpretability and performance. XGBoost outperformed other models with an AUC exceeding 98% and accuracy around 95%. Moreover, external validation confirmed the model's generalizability across heterogeneous clinical populations, suggesting that interpretable ML frameworks can support reliable and non-

invasive PCOS diagnosis.

3. METHODOLOGY

3.1 Dataset Description

Our study used a dataset that is medically proven on PCOS, which was gathered from hospitals and other diagnostic centres in India. The dataset, originally compiled by Kottarathil et al., includes clinical records from a total of 541 women of reproductive age who were evaluated for suspected PCOS, comprising 356 non-PCOS and 185 PCOS cases[10]. The gathered dataset contains features that are numerical as well as categorical that represent both hormonal and menstrual parameters. It had group of women aged between 18 to 45 years.

The dataset uses attributes like Age, Height, Weight, BMI, Blood group, AMH, LH/FSH ratio, Cycle length, Marriage status, Pimples. Hair growth. Skin darkening, pregnancy, No. of follicles, etc. The PCOS status is shown by risk level (High/Moderate/Low). The dataset displayed a medium level imbalance in class which indicates patient distributions.

3.2 Data Preprocessing

Before model training, we preprocessed the data rigorously to maintain data quality and remove inconsistency.

- Handling missing values : The numeric values that were missing were imputed by taking the mean of individual features and we handled the categorical missing values by imputing mode.
- Feature Encoding : We took the categorical variables and converted them into numeric display using label encoding.
- Handling Outliers : We capped the hormonal outliers that were extreme such as LH,FSH,AMH based on interquartile range that prevented skewed model learning.
- Feature Scaling : We kept raw values to preserve interpretability.

3.3 Correlation Analysis

In order to understand interrelationships between the variables in a better way, we computed and visualized a correlation matrix using Seaborn heatmap as shown in Fig. 1. The heatmap highlights positive as well as negative relationship among the PCOS dataset. It showed strong correlation between LH/FSH ratio and BMI, Insulin levels. These relations show hormonal imbalance and metabolic irregularities that associate to PCOS.

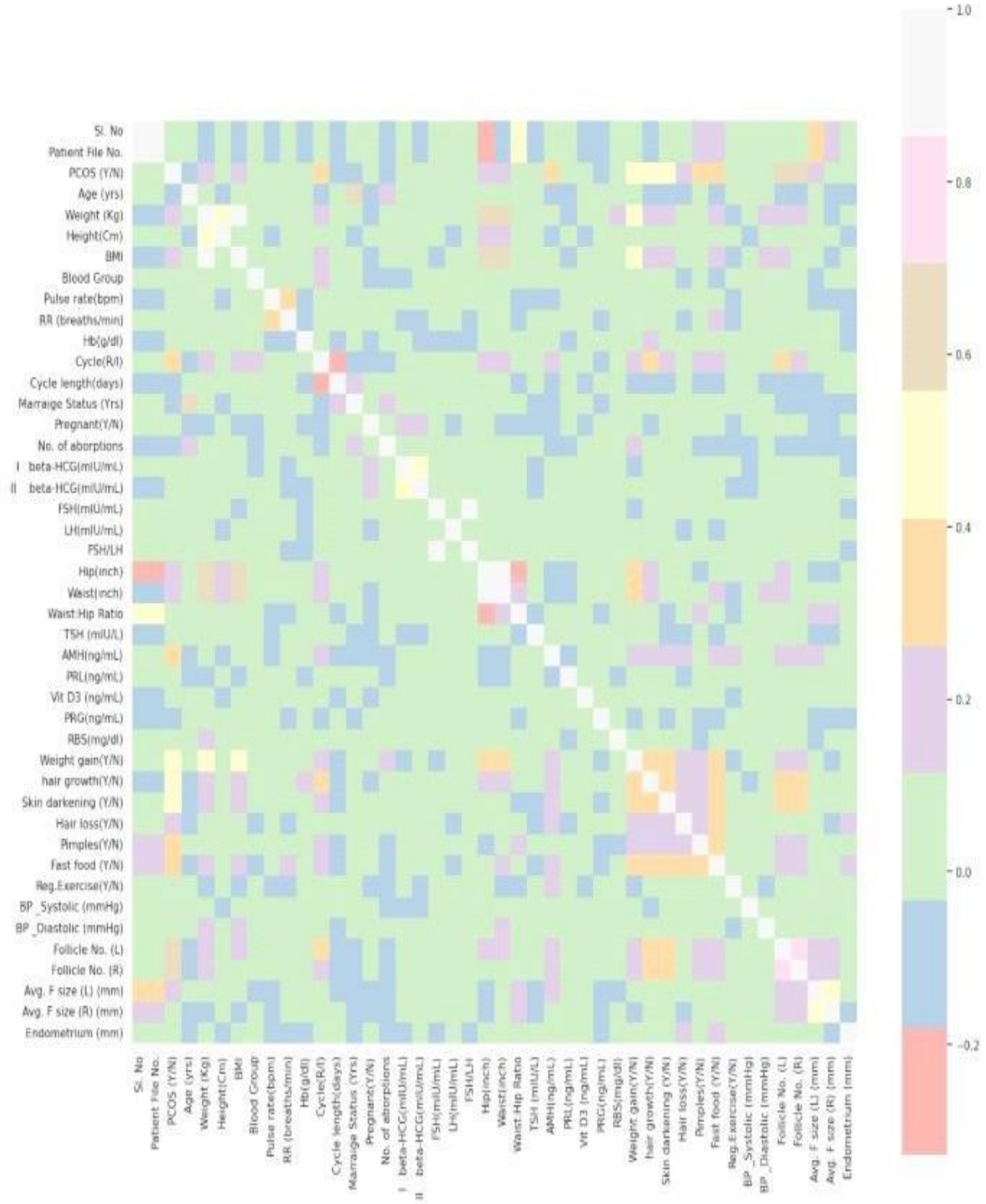

Fig. 1 : Correlation heatmap showing hormonal imbalance and metabolic irregularities

3.4 Target Correlation Analysis

In order to find which features have strongest linear association with PCOS (Y/N), we calculated correlation values between independent variable and target variable. The features that showed higher positive correlations exhibit high influence in predicting PCOS. Higher negative correlations show inverse results.

The result is given in Table 1 below. It displays the correlation with each feature of the target variable.

	PCOS (Y/N)
PCOS (Y/N)	1.000000
Follicle No. (R)	0.648327
Follicle No. (L)	0.601344
Skin darkening (Y/N)	0.471733
hair growth(Y/N)	0.454667
Weight gain(Y/N)	0.441067
Cycle(R/I)	0.401644
Fast food (Y/N)	0.370160
Pimples(Y/N)	0.286077
AMH(ng/mL)	0.264141
Weight (Kg)	0.211918
BMI	0.199534
Hair loss(Y/N)	0.172079
Waist(inch)	0.164598
Hip(inch)	0.152297
Avg. F size (L) (mm)	0.132992
Endometrium (mm)	0.106618
Avg. F size (R) (mm)	0.097698
Pulse rate(bpm)	0.091021
Hb(g/dl)	0.087179
Vit D3 (ng/mL)	0.085494
Height(Cm)	0.068254
Reg.Exercise(Y/N)	0.065317
LH(mIU/mL)	0.063879
Sl. No	0.060998
Patient File No.	0.060998
RBS(mg/dl)	0.040922
BP _Diastolic (mmHg)	0.036032
RR (breaths/min)	0.030928
Blood Group	0.030433
II beta-HCG(mIU/mL)	0.012768
Waist:Hip Ratio	0.012386
BP _Systolic (mmHg)	0.007942
PRL(ng/mL)	0.005143
TSH (mIU/L)	-0.010148
FSH/LH	-0.018316
Pregnant(Y/N)	-0.027565
I beta-HCG(mIU/mL)	-0.027617
FSH(mIU/mL)	-0.030219
PRG(ng/mL)	-0.043834
No. of aborptions	-0.057158
Marraige Status (Yrs)	-0.110058
Age (yrs)	-0.160513
Cycle length(days)	-0.176168

Table 1 : Correlation features with individual features of PCOS

3.5 Model Development

As one of the most powerful ensemble learning techniques, gradient tree boosting is particularly well-suited for tackling high-dimensional and imbalanced clinical datasets [11]. XGBoost, a highly optimized and scalable implementation of gradient boosting, offers significant advantages over alternative classifiers, including rapid computation due to advanced parallelization strategies and memory-efficient block structures[12].

The model selected was XGBoost because of its robustness in handling nonlinear feature imbalance and missing data. At start, a base model was trained for default hyperparameters. Using these default hyperparameters, we performed grid search with cross validation.

3.6 Feature Engineering

Since the dataset had high-dimensional and complex nature, it was important to select effective features that would optimize model performance. A correlation analysis was conducted to eliminate redundant and repetitive variables. All the derived features such as LH/FSH ratio, Cycle regularity were used to capture the complex relationship between physical and endocrine parameters. The values that had high correlation were cut down to enhance generalization. This action ensured that the model was non-redundant.

4. CONCLUSION

This study showcases the potential of ML specifically the XGBoost model in advancing the PCOS diagnosis through the integration of data science techniques and ML technique. By integrating this tool in clinical practices health care providers can achieve precise and efficient diagnosis based on patient symptoms and test results. Early detection by a smart predictor could enhance reproductive of thousand of women all around the globe. Our study shows that number of follicles on both ovary, average size of follicles, cycle length, cycle regularity, skin darkening, weight gain, hair growth are the attributes linked to producing a reliable PCOS diagnosis. The accuracy of the PCOS Diagnosis statues using XGBoost ML technique is 87.11.

Though a significant amount of research has been done on PCOS prediction model, the most major problem that need to be resolved is to bridge the gap between research and clinical application. Tis include developing a user friendly tool compatible with the prediction model.

Our work contributes in development of scalable, cost effective and non invasive tools for PCOS diagnosis.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our guide, **Dr. S. V. Athawale**, for his continuous support, valuable suggestions, and encouragement throughout our project, *"A machine learning approach for detecting Polycystic Ovary Syndrome using XGBoost"*

We also thank the **Department of Computer Engineering** at **A.I.S.S.M.S College of Engineering** for providing us with the required resources and guidance to carry out this work.

Finally, we are thankful to our friends and family for their constant motivation and support during the project.

REFERENCES

[1] Bozdag, G., Mumusoglu, S., Zengin, D., Karabulut, E., & Yildiz, B. O. (2016). The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. *Human Reproduction*, 31(12), 2841–2855.

[2] Azziz, R., et al. (2016). Polycystic ovary syndrome. *Nature Reviews Disease Primers*, 2(1), 16057.

[3] Stein, I. F., & Leventhal, M. L. (1935). Amenorrhea associated with bilateral polycystic ovaries. *American Journal of Obstetrics and Gynecology*, 29(2), 181–191.

[4] Norman, R. J., Dewailly, D., Legro, R. S., & Hickey, T. E. (2007). Polycystic ovary syndrome. *The Lancet*, 370(9588), 685–697.

[5] Dunaif, A. (1997). Insulin resistance and the polycystic ovary syndrome: mechanism and implications for pathogenesis. *Endocrine Reviews*, 18(6), 774–800.

[6] Diamanti-Kandarakis, E., & Dunaif, A. (2012). Insulin resistance and the polycystic ovary syndrome revisited: an update on mechanisms and implications. *Endocrine Reviews*, 33(6), 981–1030.

[7] Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertility and Sterility*, 81(1), 19–25.

[8] Patel, S., & Desai, P. (2021). Artificial Intelligence and Machine Learning in the Diagnosis of PCOS: A Review. *Journal of Biomedical Informatics*, 122, 103890.

[9] Lee, H., et al. (2020). Molecular and genetic advances in polycystic ovary syndrome. *Trends in Endocrinology & Metabolism*, 31(10), 757–770.

[10] T. Kottarathil, "PCOS (Polycystic Ovary Syndrome) Dataset," *Kaggle*,2020.[Online].Available: https://www.kaggle.com/datasets/tilakraj/pcos-dataset

[11] Jerome, H. & Friedman Greedy function approximation: A gradient boosting machine. Ann. Statist. 29, 1189–1232 (2001).

[12] M. Agirsoy and M. A. Oehlschlaeger, "A machine learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features," *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 3, pp. 5237–5249, 2023, doi: 10.3233/JIFS-230473.