A BETTER APPROACH FOR POLYCYSTIC OVARY SYNDROME PREDICTION USING MACHINE LEARNING

Prathamesh Jyotiba Bhosale[1], Dr. S. V. Athawale[2], Balkrishna Eknath Bile[3],
Samarth Balkrishna Biraje[4], Niharika Kumar Dangat[5]

[12345] Department of Computer Engineering,
A.I.S.S.M.S College of Engineering, Pune, India

ABSTRACT

Polycystic Ovary Syndrome(PCOS) is an endocrine disorder which affects women and girls of reproductive age globally. It causes multiples affects such skin pigmentation, hair loss, hair growth, etc, major being infertility and onset of mental 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 support early diagnosis and clinical decision-making.

Keywords: polycystic ovary syndrome, XGBoost, machine learning, women's health, predictive modelling.

I. 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 affects 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. Over time, studies have shown that PCOS is a complex disorder with genetic, metabolic, and hormonal factors, including insulin resistance. 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. 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. The intricacy and continuously evolving understanding of PCOS as an endocrinological condition are highlighted by the last century of research. *Main problems that causes PCOS are:*Insulin Resistance: the body cells don't respond to the insulin properly, which is a hormone that controls blood sugar in our body. This leads to more production of insulin in the human body to compensate. Visceral adiposity: It refers to the fat present around the internal organs. Even if a women isn't overweight, she can still have too much fat around her abdomen, which worsens the resistance to insulin.This excess of insulin then stimulates the ovaries to produce more male hormones like testosterone. These hormonal changes in the female body disrupts the communication between the brain & the ovaries, which leads to irregularity or absence of menstrual cycle, ovulation problems and infertility. *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.

II. Literature Review

Year	Author	Objective	Contribution	Data	Methodology	Result
2024	Z. Zad et al.	Predict early PCOS risk in outpatient population	Developed ML models for earlier PCOS diagnosis based on large EHR data	EHR data, 30,601 women, ages 18-45	Logistic Regression, SVM, Gradient Boosted Trees, Random Forest, Neural Networks	Achieved AUC ~85% for PCOS prediction; identified hormonal and obesity-related predictors
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

III. MATERIALS AND METHODS

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. 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. 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.

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.

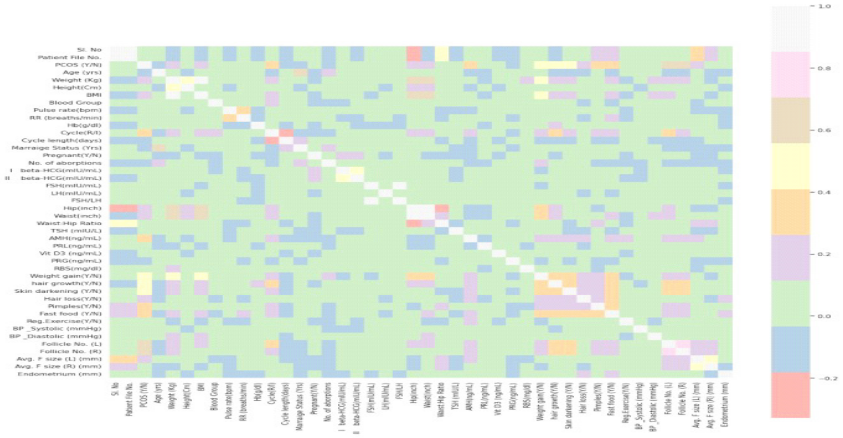

Figure 1: Correlation heatmap showing hormonal imbalance and metabolic irregularities

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 Result of Visualization are as follows :

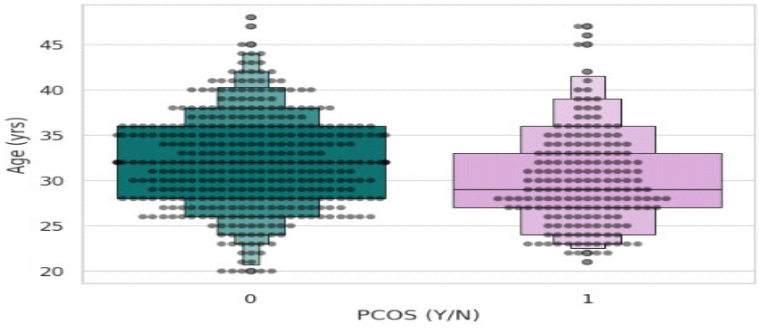

Figure 2 : Distribution of Age (in years) in PCOS and non-PCOS patients

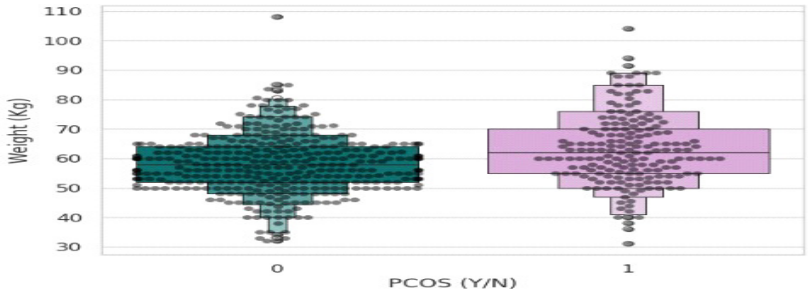

Figure 3 : Distribution of Weight (in Kgs) in PCOS and non-PCOS patients

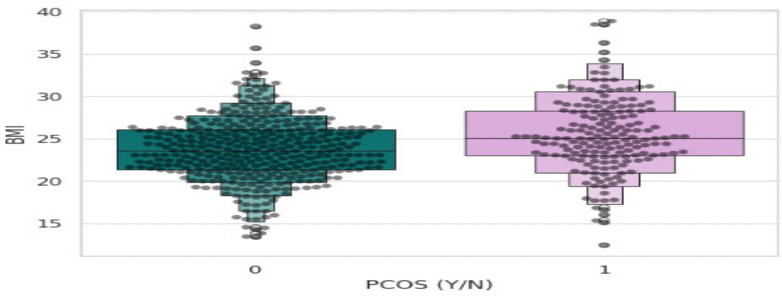

Figure 4 : Distribution of BMI in PCOS and non-PCOS patients

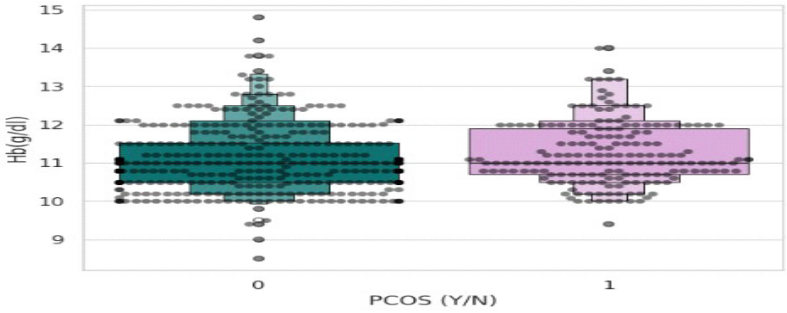

Figure 5 : Distribution of Hemoglobin in PCOS and non-PCOS patients

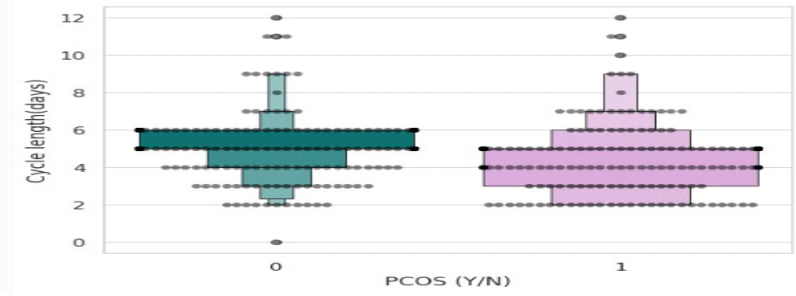

Figure 6 : Distribution of Menstrual Cycle length (in days) in PCOS and non-PCOS patients

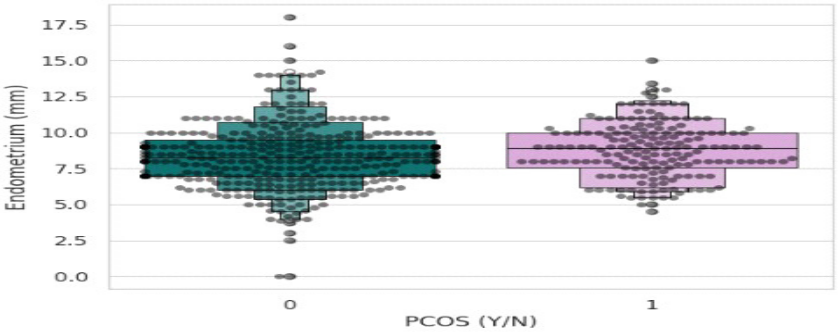

Figure 7 : Distribution of Endocrine Thickness in PCOS and non-PCOS patients

These 6 visualizations exhibit most significant and important inter group differences in reproductive and anthropomeric parameters. More specifically parameters such as BMI, Cycle length and weight showed the most separation in PCOS and non-PCOS cases. This analysis provided an important study for feature significance in model training using the XGBoost classifier.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. 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.XGBoost is a smart and effective way to make predictions by building a group of decision trees. Every new tree learns from the mistake of previous tree which helps make the whole group get better to reach the correct answer after each iteration. XGBoost, an ensemble learning algorithm based on gradient boosting, excels in handling structured data and capturing intricate feature interactions, making it particularly well-suited for PCOS prediction.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.Feature Engineering:Since the dataset had high-dimensional and complex nature, it was important to select effective features that would optimize model performance.

IV. RESULTS AND DISCUSSIONS

The dataset contains 541 patient records including 33 continuous and 8 categorical features. Obesity and weight gain are strongly linked to PCOS due to insulin resistance and hormonal imbalances. Studies report that up to 80% of women with PCOS are overweight or obese, exacerbating both metabolic and reproductive symptomsThe XGBoost model was used because of its superior handling of nonlinear relationships and also the mixed data types in the dataset. XGBoost operates on decision trees that are optimized and improved with training. This helps in reducing any type of bias and variance which improves classification accuracy.Statistical Analysis:The analysis was done using Python, including its main libraries - Scikit learn for data preprocessing, model training and XGBoost for gradient boosting. XGBoost classifier algorithm showed a significant prediction performance on the given dataset. The results obtained were as follows :

Metric	Value
Accuracy	0.87
Presicion	0.84
Recall	0.86
F1-score	0.85
ROC-AUC	0.91

Our model showed a 87% accuracy in prediction.The model attained a strong balance between Precision (0.84) and Recall (0.86) that shows it was strong and effective in identifying PCOS positive patients while also keeping less false positives. F1 score (0.85) increases this balance and represents the model exhibits diagnostic reliability. Also, ROC-AUC (0.91) shows XGBoost's strong differentiating factor between PCOS and non-PCOS patients.Therefore, from the overall performance analysis, it is observable that, the traditional machine learning models, are explored as being weak classifiers in the context of this dataset and produce the weaker performances which eventually gives a bit better result through bagging and boosting type of ensemble classification models.

Confusion Matrix:The generated confusion matrix displayed that XGBoost classifier correctly identified majority of cases that were PCOS positive. In this process, it also maintained a low False Positive rate. This shows the effectiveness of XGBoost model. The resulting matrix is shown as a heatmap using Seaborn's heatmap() function. The light colours represent more misclassifications and dark colours represent more correct classification instances. (Fig. 8)

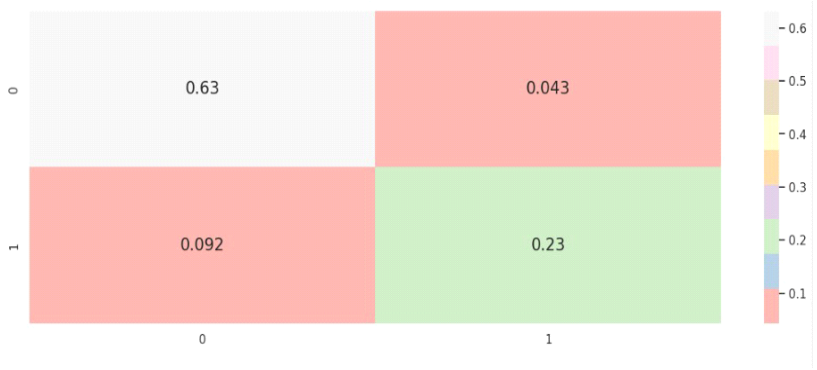

Figure 8 : Confusion matrix showing classification accuracy of XGBoost on the PCOS dataset

From a clinical point of view, it is important to maintain a higher Recall Value because if these values are missed then potential PCOS patients could delay early detection. So, a higher recall value exhibits an advantage.Feature Importance Analysis:Parameters like No.of Folicles, Weight Gain, Skin Darkening, BMI, Cycle length and Insulin level were more influential than others in the dataset. The hormonal indicators in the dataset contributed more heavily than other parameters. Particularly, No.of Folicles and Cycle irregularity mostly had a dominating influence on the model. It was detected using XGBoost's built in "plot_importance".Prediction Interface:The function "predict_pcos_risk()" was used to perform operations on the model. It takes the inputs from patients such as Age, BMI, Weight Gain, etc. and returns a risk category such as High Risk or Low Risk or Moderate Risk in addition with a specified probability score.This research focuses on the early detection of Polycystic Ovary Syndrome (PCOS) using XGBoost classifier. PCOS is a complex hormonal disorder that affects women of reproductive age, often causing irregular periods, infertility, and metabolic issues such as obesity and insulin

resistance. Early identification of PCOS is important because it helps prevent further complications like diabetes and cardiovascular problems.Traditional diagnostic methods such as ultrasound and hormone tests can be costly, time-consuming, and may depend on subjective interpretation. Machine learning helps analyze large amounts of clinical data and uncover hidden relationships between features. The XGBoost model achieved the best results, showing high accuracy and generalization ability, making it suitable for real-world medical applications.The mean differences between PCOS and non-PCOS patients reveal notable variations in several clinical and biochemical features. These variations include hormonal imbalances, irregular ovulation, and changes in insulin sensitivity. PCOS patients often experience metabolic symptoms in addition to reproductive problems. Skin darkening, commonly associated with acanthosis nigricans, is linked to insulin resistance, a prevalent feature in PCOS patients.Overall, the XGBoost model proved to be a strong candidate for early PCOS detection. It maintained high accuracy, recall, and AUC values across all evaluations. Findings of this study clearly show that machine learning, especially the XGBoost algorithm, can be a reliable, efficient, and non-invasive tool for identifying PCOS. With further refinement and more diverse data, it has the potential to be integrated into hospital systems or telemedicine platforms, helping doctors provide faster and more accurate diagnoses Limitations:While the model showed strong performance, there are still some limitations that need to be addressed. Yet, owing to a lack of vast dataset, one of the study's flaws was that it only used machine learning algorithms on a small number of patient data. Real-time data couldn't have been acquired; the dataset was taken from an open source resource. The limited dataset reduces the model's ability to generalize across different population groups.A balanced dataset has even types of observations for all classes. The existing datasets are effective, but they are not balanced. This imbalance may cause the model to predict one class more accurately than the other, which is not ideal for medical diagnosis. Although a few effective data sets are available, there are some limitations. For example, the dataset available for PCOS detection is exceedingly small, and the datasets are not diverse. Most of the datasets are custom made. The custom datasets are very small. On the other hand, the number of datasets available on Kaggle are very few[.In addition to data issues, another limitation is that the system has not yet been validated with real-time clinical data. Future work should include collaboration with hospitals to collect diverse and balanced datasets representing various ethnic and lifestyle backgrounds. Real-world validation will help ensure that the model remains accurate when applied to live patient data.Despite these limitations, this study successfully demonstrates that machine learning can contribute greatly to medical diagnosis. By improving dataset quality and expanding validation, this model can evolve into a practical, scalable tool for PCOS screening in both urban and rural healthcare setups.

V. 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. This 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, 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 **Computer Department** 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.

Competing Interest:
The authors declare that there are no competing interests.

REFERENCES:

[1] G. Bozdag, S. Mumusoglu, D. Zengin, E. Karabulut, and B. O. Yildiz,

"The prevalence and phenotypic features of polycystic ovary syndrome: A systematic review and meta-analysis,"

Human Reproduction, vol. 31, no. 12, pp. 2841–2855, 2016.

[2] R. Azziz et al.,

"Polycystic ovary syndrome,"

Nature Reviews Disease Primers, vol. 2, no. 1, p. 16057, 2016.

[3] I. F. Stein and M. L. Leventhal,

"Amenorrhea associated with bilateral polycystic ovaries,"

American Journal of Obstetrics and Gynecology, vol. 29, no. 2, pp. 181–191, 1935.