EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

# Comparative Analysis of Polycystic Ovary Syndrome Detection Using Machine Learning Algorithms

Neha Yadav[1], Ranjith Kumar A.[2] and Sagar Dhanraj Pande[3, *]

[1,2] School of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India
[3]School of Engineering and Technology, Pimpri Chinchwad University (PCU), Pune, Maharashtra, India

INTRODUCTION: Polycystic Ovary Syndrome is a condition in which the ovaries manufacture androgen, seen in small traces, resulting in the production of cysts. Menstrual cycle abnormalities, clinical and/or biochemical hyperandrogenism, and the presence of polycystic ovaries on ultrasound should all be used to diagnose PCOS. PCOS appears to be a multifaceted illness influenced by both genetic and environmental factors and the symptoms include excessive hair on the face and body, weight gain, voice changes, skin type changes, and irregular periods.
OBJECTIVES: This is the objective of this paper is to identify PCOS in its initial stage.
METHODS: To address this issue the study proposes a comparison of various machine learning algorithms and optimization techniques Among which GSCV gave the best result of 94% accuracy, followed by TPOT with 91% accuracy. Additionally, we also applied Feature selection methods to eliminate zero-importance features to increase the accuracy of algorithms.
RESULTS: The main results obtained in this paper This study explored various Feature selection techniques, ML and DL models. It is shown that Grid Search CV and TPOT classifier were best classifiers with 94% and 91% respectively.
CONCLUSION: These are the conclusions of this paper and this study will explore various DL methodologies and try to find out best optimal results for the PCOS Detection. And also, to develop an PCOS detection application to keep track of menstrual cycles and track activities and symptoms for PCOS.

**Keywords:** Polycystic Ovary Syndrome, RSCV, GSCV, BO, Optuna, TPOT

Received on 18 December 2023, accepted on 18 March 2024, published on 26 March 2024



doi: 10.4108/eetpht.10.5552

*Corresponding author. Email: sagarpande30@gmail.com

## 1. Introduction

Polycystic Ovary Syndrome is a prevalent disease that affects women before menopause stage. It is differentiated by hormonal irregularities, irregular menstrual periods, and physical metabolic problems. PCOS can cause a variety of health difficulties, including infertility, obesity, insulin resistance and cardiovascular problems. PCOS screening and diagnosis are critical for optimal management and reduction of associated health concerns. It is crucial to understand that PCOS is a multifaceted condition with varied combinations of causes affecting different persons. Some of the Key factors that affect PCOS include Genetics, Hormonal imbalance, Insulin Resistance, Inflammation and Abnormal Follicle Development.
According to World Health Organization (WHO) [1], Polycystic ovary syndrome is most common hormonal disorders among women of age between 12-55, accounting for between 8% and 13% of all reproductive-age women. Up to 70% of PCOS cases go untreated. Polycystic ovarian syndrome symptoms vary from person to person. Symptoms might alter over time and frequently arise without a clear cause. Anxiety, despair, and a negative body image can all result from PCOS. Some symptoms, such as infertility, obesity, and excessive hair growth, may result in societal stigma. This can have an impact on other aspects of one's life, such as family, relationships, employment, and community activity.
Recently in ENDO 2023[2] research shows women with polycystic ovarian syndrome (PCOS) have more body image issues than those without the illness. PCOS has no broadly accepted definition. The 2003 Rotterdam criteria, which extended on the 1990 National Institutes of Health (NIH) definition, are a well-known classification. According to the Rotterdam [3] definition, the syndrome must contain among these three criteria: hyperandrogenism (clinical or biochemical), ovulatory failure, or PCOM.As previously stated, this enlarged the 1990 NIH criteria, which classified PCOS as the presence of ovulatory failure

with hyperandrogenism but did not include ultrasonographic evidence.

## 2. Literature Survey

A thorough review of papers on PCOS and systems to assist its diagnosis was conducted and table 1 present comparative analysis of existing mythologies and papers. Amsy Denny et.al [4] proposed a method to detect PCOS in which the most suitable algorithm was RFC with accuracy of 89.02%. According to authors of this paper [5], results shows that RFLR gives 91.01 % accuracy and of 90% recall using CS (cross validation)

In study proposed by [6] Preeti Chauhan Et.al, DT classifier gives out most accurate results. In another such paper the authors used Bayesian and Logistic Classifiers in their study, and the comparison showed that BC with 93.93% accuracy outperformed LC with 91.04% accuracy [7]. Anuradha et al. [8] used Python to identify PCOS using three machine learning classification methods such as ANN, KNN, and LR on a dataset of 84 occurrences and 13 characteristics. According to the findings of this study, the most significant symptoms include acne, irregular periods, LH, sonography, and weight. Linear regression has the highest accuracy (100%), followed by ANN (94%). Authors K. Meena Et.al conducted a study on attribute reduction and found that combining feature selection and classification produces the best results for PCOS prediction [9].

The present study attempts to address gaps in PCOS Detection. First the study explored several methodologies and numerous features in which most prominent features were found out by Feature Selection methods such as using LBM. Second to Fine Tune the model we applied numerous optimization techniques and compared them on their results.

Table 1. Comparative analysis of various papers of PCOS detection

| Citation Number | Methodology | Advantages | Limitations |
|---|---|---|---|
| [10] | KNN and LR | F1 score of LR is 0.92 % | Not explored various machine learning algorithms. |
| [4] | PCA is used for ML techniques such as KNN, NB, CART, SVM, RFC | RFC achieved results of 89.02% | Didn't combine USG and biochemical result. |
| [11] | used Rapid Miner instead of Scikit Learn package | Selected 10 features using Feature Selecting techniques and achieved 93.12% accuracy using RapidMiner. | - |
| [5] | Hybrid model using RFLR | RFLR exhibits 88.52% accuracy using cross validation on prominent features | Accuracy can be optimized more |
| [7] | BC and LRC were used | PCOS labels are segmented using t test | Only comparison of two classifiers |

## 3. Methodology

### 3.1. Dataset Description

We have used dataset consisting of 512 rows which is publicly available on Kaggle. This dataset consists of 362 rows with people not having PCOS (Label-0) and 176 people having PCOS (Label -1) and displayed in Figure 1. To balance out the data we have used SMOTE over training data, and we got 430 rows with 0 and 1 respectively (Figure 2).

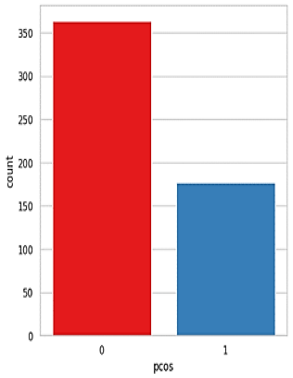


**Figure 1.** Dataset before SMOTE

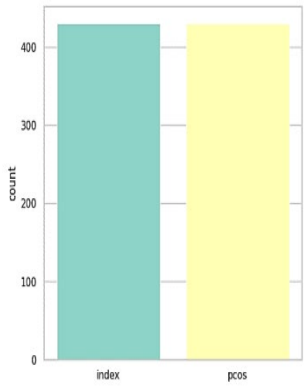


**Figure 2.** Dataset after applying SMOTE.

## 3.2. Data Preprocessing

### Feature Selection

We used LGM for Feature selection. It distinguishes the features with zero or no importance according to a GBM. The LGM is trained with early stopping using a validation set to prevent overfitting. Kernel density estimation (KDE) of different features is plotted is plotted in Figure 3. Proposed Methodology is depicted in Figure 4.

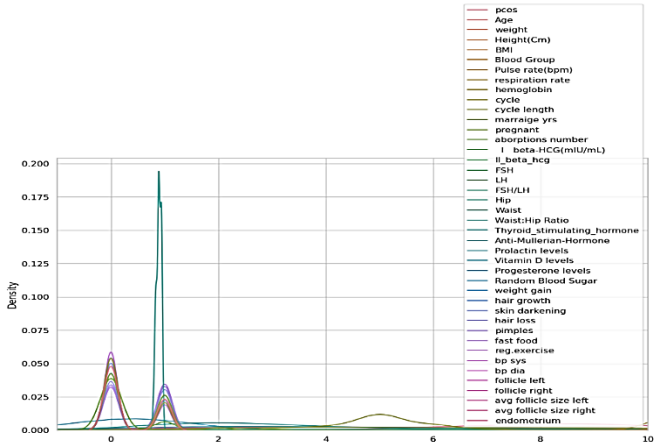


**Figure 3.** Kernal density estimation of various features in PCOS

## 3.3. Proposed Methodology

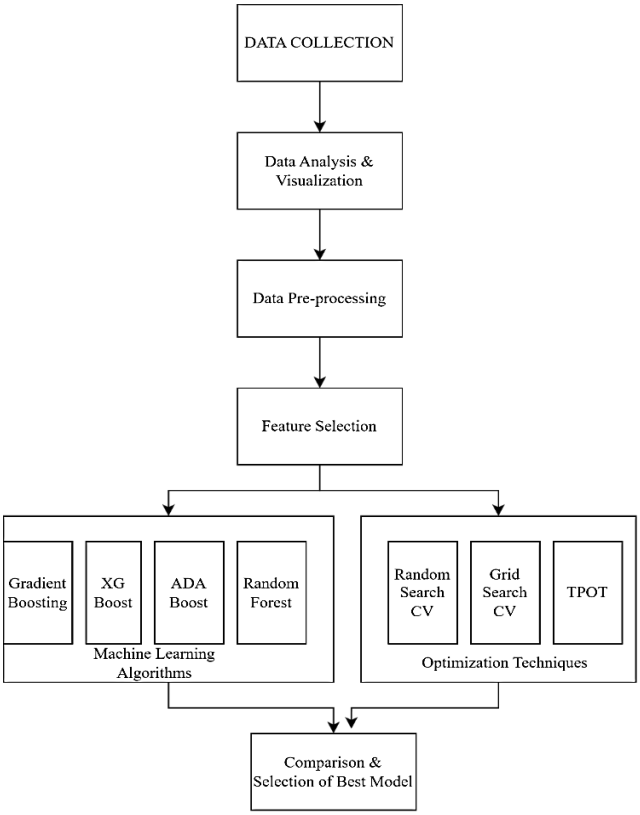


**Figure 4.** Proposed Methodology

## 3.4. Optimization Techniques

Optimization is used to decrease and increase a specific a specific objective function i.e. it is used to identify the optimum possible values of parameters for a particular model.

***Randomized Search CV*** Optimization is used to decrease and increase a specific a specific objective function i.e. it is used to identify the optimum possible values of parameters for a particular model.

***Grid Search CV*** Optimization is used to decrease and increase a specific a specific objective function i.e. it is used to identify the optimum possible values of parameters for a particular model.

***Bayesian Optimization*** a function's minimum is discovered using probability. Finding the input value to a function that can produce the lowest possible output value is the goal. It typically outperforms random, grid, and manual search, offering greater performance during the testing phase and requiring less time for optimisation.

***Optuna*** Optuna is well-suited for discovering the best hyperparameters even in high-dimensional and complicated search spaces since it uses Bayesian optimisation and other cutting-edge techniques to intelligently explore the hyperparameter space.

## 4. Results

The evaluation measures for machine learning algorithms are measured via Accu, Pre, Reca and F1 scores, as well as CM and ROCs, which are defined as:

***Accu (Accuracy):*** Accu is a typical statistic used to quantify the overall correctness of a classification model. It estimates the ratio of properly predicted instances to the total instances.

$$Accu = \frac{TP + TN}{TP + TN + FP + FN}$$

***Pre (Precision):*** Pre is the percentage of accurately predicted positive instances (TP predictions) among all instances that are projected to be positive.

$$Pre = \frac{TP}{TP + FP}$$

***Reca (R):*** The percentage of samples the models correctly identify among all the positive samples.

$$Reca = \frac{TP}{TP + FN}$$

***F1:*** It is measure or metric used to evaluate performance of model.

$$F1 = \frac{2 * Pre * Reca}{Pre + Reca}$$

We perform comparative analysis of various algorithms to see which one performs better and Table 2 describes the comparison of various ML algorithms specially Ensemble Learning algorithms.

Table 2. Comparative analysis of various papers of PCOS detection

| Algorithms | Labels | Pre | Rec | F1 | Accu |
|---|---|---|---|---|---|
| Gradient Boosting | 0 | 0.92 | 0.91 | 0.91 | 0.88 |
| | 1 | 0.79 | 0.82 | 0.81 | |
| XG Boost | 0 | 0.90 | 0.88 | 0.89 | 0.85 |
| | 1 | 0.74 | 0.79 | 0.76 | |
| Ada Boost | 0 | 0.88 | 0.91 | 0.89 | 0.75 |
| | 1 | 0.77 | 0.73 | 0.75 | |
| Random Forest | 0 | 0.91 | 0.92 | 0.91 | 0.88 |
| | 1 | 0.81 | 0.79 | 0.80 | |

We also applied various optimization techniques to check the performance of the models and fine-tuning them with different parameters. Table 3 describes analysis of Optimization Techniques.
The results emphasise that supplementing the model with discriminant characteristics produces better outcomes after completing numerous experiments with adjustments in the utilised features. In addition, despite its long use, the Grid Search CV has the best outcomes and invariant behaviour. intricacy of performance with accuracy of 94% followed by TPOT classifier with 0.91%. Abbreviations used in this paper are given in Table 4.

Table 3. Comparative analysis of various Optimization algorithms

| Optimization Algorithms | Labels | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|---|
| Random Search CV | 0 | 0.90 | 0.93 | 0.92 | 0.88 |
| | 1 | 0.83 | 0.76 | 0.79 | |
| Grid Search CV | 0 | 0.90 | 0.95 | 0.92 | 0.94 |
| | 1 | 0.86 | 0.76 | 0.81 | |
| TPOT | 0 | 0.91 | 0.92 | 0.93 | 0.91 |
| | 1 | 0.85 | 0.79 | 0.74 | |

Table 4. Abbreviations List.

| Abbreviation | Phrase |
|---|---|
| SMOTE | Synthetic Minority Oversampling Technique |
| LGBM | Light Gradient Boosting Machine |
| Accu | Accuracy |

| | |
|---|---|
| Pre | Precision |
| Reca | Recall |
| CM | Confusion Matrix |
| TP | True Positive |
| TN | True Negative |
| FP | False Positive |
| FN | False Negative |
| RSCV | Randomized Search CV |
| GSCV | Grid Search CV |
| BO | Bayesian Optimization |
| PCA | Priniciple Component Analysis |
| KNN | K- Nearest Neighbour |
| RFLR | Random Forest and Linear Regression Classifier |
| GBM | Gradient Boosting Machine |
| RFC | Random Forest Classifier |
| LRC | Linear Regression Classifier |
| BC | Bayesian Classifier |

## 5. Conclusion and Future Scope

PCOS occurs when the body creates excess hormones. This study explored various Feature selection techniques, ML and DL models. It is shown that Grid Search CV and TPOT classifier were best classifiers with 94% and 91% respectively.

Furthermore, the study will explore various DL methodologies and try to find out best optimal results for the PCOS Detection. And also, to develop an PCOS detection application to keep track of menstrual cycles and track activities and symptoms for PCOS.

**System Configuration**

Operating System: Windows, Mac, Linux

SDK: OpenCV, TensorFlow, Keras, Numpy

The Hardware interfaces Required are:

Camera: Good quality 2MP

Ram: Minimum 4GB or higher

Processor: Intel or Amd Ryzen

## References

[1] **Article:** WHO:https://www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome

[2] **Journal Article:** ENDO 2023: Chicago IL [2023]

https://www.contemporaryobgyn.net/view/getting-a-grip-on-polycystic-ovary-syndrome

[3] **Journal Article:** Getting a grip on polycystic ovary syndrome: Jessica L.Chan, MD [2023]:

https://www.contemporaryobgyn.net/view/getting-a-grip-on-polycystic-ovary-syndrome

[4] **Conference:** Amsy Denny, Anita Raj and Ashi Ashok : "i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques".

[5] **Conference:** Subrato Bharati, Prajoy Podder, M. Rubaiyat Hossain Mondal: "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms"

[6] **Conference:** Preeti Chaunhan, Pooja Patil, Neha Rane: "Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS".

[7] **Conference:** P. Mehrotra et al. "Automated screening of Polycystic Ovary Syndrome using machine learning techniques".

[8] **Journal:** Anuradha, D.T., & Priyanka R. L., Genetic Clustering for Polycystic Ovary Syndrome Detection in Women of Reproductive Age

[9] **Journal:** K. Meena, M. Manimekalai, and S. Rethinavalli. "Correlation of Artificial Neural Network Classification and NFRS attribute Filtering algorithm for PCOS data"

[10] **Journal Article:** Namrata Tanwani : "Detecting PCOS using Machine Learning"

[11] **Journal:** Satish C. R Nandipati, Chew XinYing and Khaw Khai Wah: Polycystic Ovarian Syndrome (PCOS) Classification and Feature Selection by Machine Learning Techniques