Polycystic Ovary Syndrome(PCOS) Classification

Hriday Pradhan
1910110163
Shiv Nadar University
hp103@snu.edu.in

Mehak Agrawal
1910110233
Shiv Nadar University
ma286@snu.edu.in

Shraddha Arora
1910110375
Shiv Nadar University
sa350@snu.edu.in

***Abstract*—Polycystic ovary syndrome (PCOS) is the most common endocrine disorder affecting many women in their child-bearing age groups. This disturbance results in problems affecting many body systems like uneven mensuration cycle, obesity, oily skin, pimples, anxiety disorders, etc. Women having PCOS have irregularity in menstrual periods as well as cyst formations in either or both ovaries. Symptoms of PCOS are: irregular periods, excess androgen, polycystic ovaries, abnormal BMI, disturbed levels of hormones (LH, FSH, DHEAS), and poor insulin resistance. If not diagnosed in time, the condition can cause serious health issues. To overcome this problem, this paper proposes to develop an application for the early prediction of PCOS using Machine Learning techniques. The required dataset is cleaned using Python and Google Colab. Classification of PCOS is done using various machine learning techniques such as K-Nearest Neighbor(KNN), Decision Tree Classifier, Support Vector Machine(SVM), Logistic Regression(LR). Based on the accuracy and confusion matrix, the CatBoost Classifier was found to be the most accurate model for PCOS prediction.**

I. Introduction

There are numerous abnormalities affecting women's reproductive systems that could lead to major health problems in the future. These conditions affect the ovaries, uterus, cervix, vagina, and other reproductive organs. Hormonal changes in the body, hormonal imbalance, irregular living patterns, stress, and other factors all contribute to the occurrence of these diseases. Polycystic ovarian syndrome (PCOS) is a type of hyperandrogenism in which the ovaries produce too much androgen. It's a condition that affects a lot of women in their reproductive years (15-40 yrs). Hormone levels in women are impacted in this situation. Cysts grow on the ovaries' periphery as a result of the hormonal imbalance. Cysts resemble follicles or little balls of tissues and can be present in one or both ovaries in PCOS women. These cysts are quite little and are completely harmless. The size and number of these cysts aren't set in stone. They can range in size from 2mm to 9mm. Excess androgen, an irregular menstrual cycle, and polycystic ovaries are the main symptoms of PCOS. Many women with PCOS will be able to conceive with the correct medication. However, only around half of all women with PCOS are adequately identified, leaving many people misdiagnosed. To diagnose PCOS, a variety of tests must be carried out. Early detection and diagnosis of PCOS using limited tests and imaging technology is critical since the disorder causes ovarian malfunction, which increases the chance of miscarriage, sterility, or even ovarian cancer, as well as mental anguish for patients owing to time and money wasted. A doctor may conduct a physical examination, a Pelvic ultrasound test, and some blood tests to diagnose PCOS and rule out other possible reasons of your symptoms.

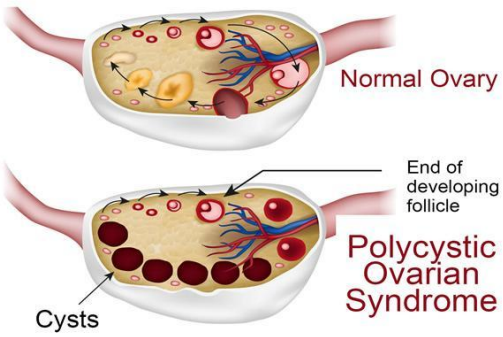


Fig. 1. Normal vs Polycystic ovarian Syndrome Ovary

II. Literature review

Priyanka R. Lele, Anuradha D. Thakare, "Comparative Analysis of Classifiers for Polycystic Ovary Syndrome Detection using Various Statistical Measures" This paper aimed to use the different Classification algorithms; Multilayer Perceptron, K star, IB1 instance-based, Locally weighted learning, Decision Table, M5 rules, Zero R, Random Forest and Random Tree algorithm to detect whether the patient has PCOS or not. The dataset for the proposed system wasn't readily found in available repositories. Therefore, the dataset was created in discussion with a medical practitioner with expertise in PCOS detection. The dataset generated had 13 attributes and 2 classes. A total of 40 instances were created. It was observed that in all statistical parameters, the K star algorithm outperformed the other algorithms, giving good classification accuracy.

Preeti Chauhan, Pooja Patil, Neha Rane, Dr Pooja Raundale, Harshil Kanakia, "Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS" This research aimed to predict PCOS in women from the symptoms provided. For this, questionnaires using Google forms was created and circulated to women of all age groups. A total of 267 responses were available for this study which was collected through a survey. The data comprised women from all age groups. Among the data collected, 206 cases were normal non-PCOS cases, while the remaining 61 cases were of women suffering from

PCOS. There were a total of 27 features. Moreover, a mobile application for early detection of PCOS using Decision Tree Classifier was implemented. The best performance was given by the Decision Tree Classifier with an accuracy of 81%. It also gave precision of 70% and specificity of 94%.

Shakoor Ahmad Bhat, Dr Rashmi Gupta, "Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms" This paper proposed a new automated system which focused to help medical expertise in PCOS detection and further helping in the treatment of PCOS patients. In this research, a novel approach was proposed that is the detection of PCOS using a hybrid machine learning model and CatBoost model. Using the AUC score and ROC curve plot, it was quite clear that XGBRF performed better than CatBoost which means it accurately distinguished the Normal women and PCOS women.

III. Proposed Model

The proposed system classifies the process from importing libraries to model selection in a flowchart.

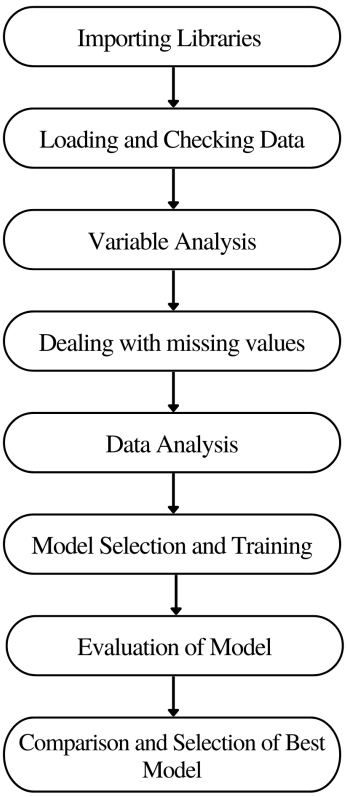


Fig. 2. Proposed Model

The first step to building our model was to import our libraries and datasets into our Google Colab notebook. Once we imported our libraries, we then went ahead to import our dataset.After opening the dataset, we saw the range of features or columns we have, from the patient's age to the thickness of their Endometrium. The files were divided into infertility and without-infertility patients so we combined them and deleted the repeated features. We also got rid of some of the columns, making the data easier to use. We checked the data types of the columns within our dataset and converted all of our data into numerical values for our model to be able to process it. Then, for the Categorical variables, we looked at the value count and converted those into values and plotted graphs of the same. Next, we dealt with the missing values by filling them up with the median value of the features. After which, we examined the correlation matrix of all features and understood the data through the data visualization technique. We had partitioned the dataset into two subsets namely the training set and the testing set. The dataset was split in the ratio of 7:3 as the train to test dataset. We then calculated the training and testing accuracy by Simple Logistic Regression and used the below-mentioned classifier algorithms:

- Decision Tree
- SVM
- Random Forest
- KNN
- Logistic Regression
- XGBRF
- CatBoost

IV. Methodology

A. Importing Libraries

The first step to building our model was to import our libraries and datasets into our Google Colab notebook. The following libraries were imported into pandas: The most popular python library that is used for data manipulation and analysis. In this project, it is primarily useful for data frame manipulation.

- numpy: A python library that provides support for large, multi-dimensional arrays and matrices, and has high-level mathematical functions to help operate on and manipulate these arrays.
- matplotlib.pyplot and seaborn: Used for data visualization.

B. Loading and Checking Data

The dataset we used for this project contains data from 541 patients across 10 hospitals in Kerala, India. It contains all physical and clinical parameters to determine PCOS and infertility related issues. Once we have imported our libraries, we can import our dataset. However, before importing our data, we need to download it and remove any unnecessary columns, making it much easier to use once it is downloaded. Once we have opened the dataset, we can see the range of features or columns we have, from the patient's age to the thickness of their Endometrium. The files are divided into infertility and without-infertility patients so we'll combine them i.e basically we'll merge them on the basis of patient file no. Further, we'll delete the repeated features and finally, we'll change this column i.e PCOS(Y/N) to Target. We can also get rid of some of these columns, making the data easier to use once we have imported it into our notebook. As we can see, the column titled "Sl. No" and "Patient File No." will not

serve any purpose as they don't need to be included within the input, so we can delete both of these columns. We want to see the datatypes of the columns within our dataset, which we can do by running data.info( ), which shows us that all of our data is of float datatype except for 2 of them which are object datatype. However, we need to convert our data into numerical values for our model to be able to process it. For this reason, we do the variable analysis of the data.

C. Variable Analysis

For the Categorical variables, we looked at the value count and converted those into values. We assigned 1 to yes and 0 to no. Similarly, we assigned values to Blood Group: A+ = 11, A- = 12, B+ = 13, B- = 14, O+ =15, O- = 16, AB+ =17, AB- = 18.

Now that we have cleaned up our data, we can visualize the distribution of data within certain features through histograms.

Categorical Variables

Fig. 3. Fast Food

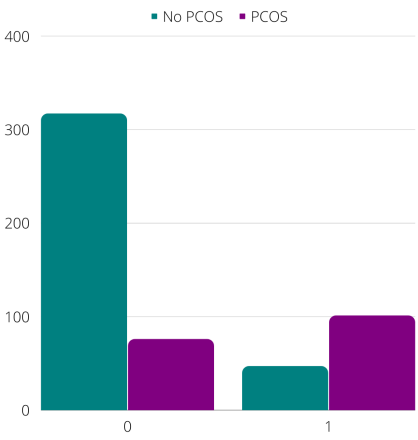


Fig. 4. Hair Growth

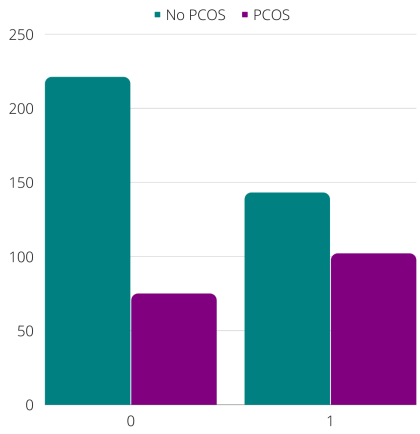


Fig. 5. Hair Loss

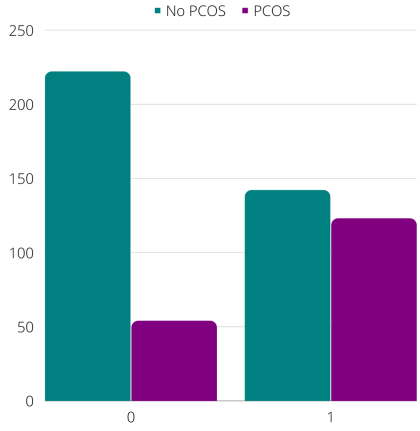


Fig. 6. Pimples

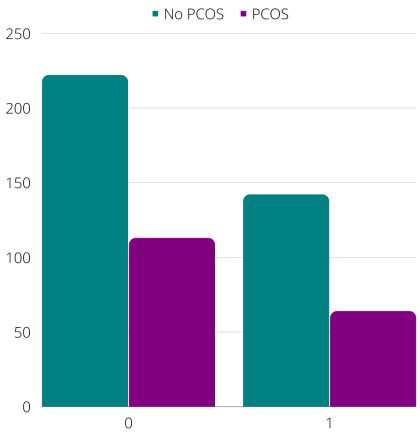


Fig. 7. Pregnant

Fig. 8. Skin Darkening

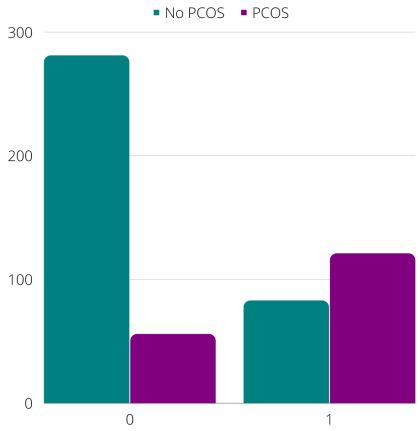


Fig. 9. Weight Gain

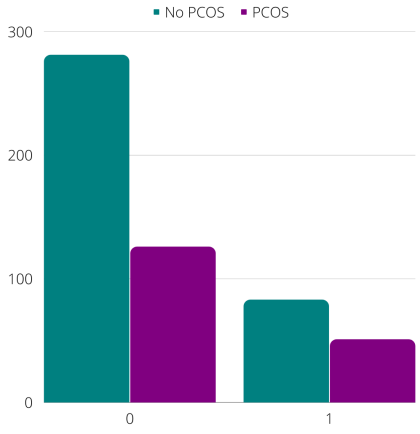


Fig. 10. Regular Exercise

Numerical Variables

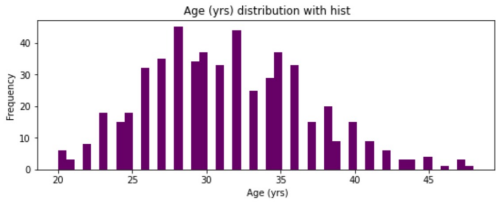


Fig. 11. Age

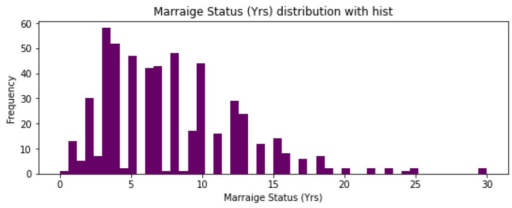


Fig. 12. Marital Status

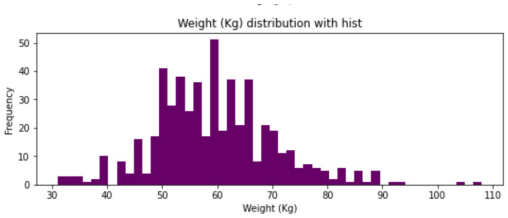


Fig. 13. Weight

D. Dealing with missing values

We then dealt with the missing values by filling them up with the median value of the features.

E. Data Analysis

We looked at the correlations within our data. First, we found the correlations between the columns, or features, of our data by running df.corr( ).
Next, we visualized these correlations by creating a heatmap. The figsize establishes the size of the heatmap, annot = True, shows us the numerical values inside of the heatmap, and fmt = '.2f' shows us the numerical correlation rounded to the hundredths place. (Fig. 14.)
We concentrated on the properties that have a relationship greater than 0.25 with the target. (Fig. 15.)

F. Model Selection and training

We split 70% of the data into training data, and 30% of it into testing data. The training data was simply used to train the model. We just fed the data to the model, so that it can learn the relationships between the inputs and the outputs. The testing data was used after the model was trained. After we had finished the training and iterations, we fed the testing data to the model. The model will never have seen the testing data during training.

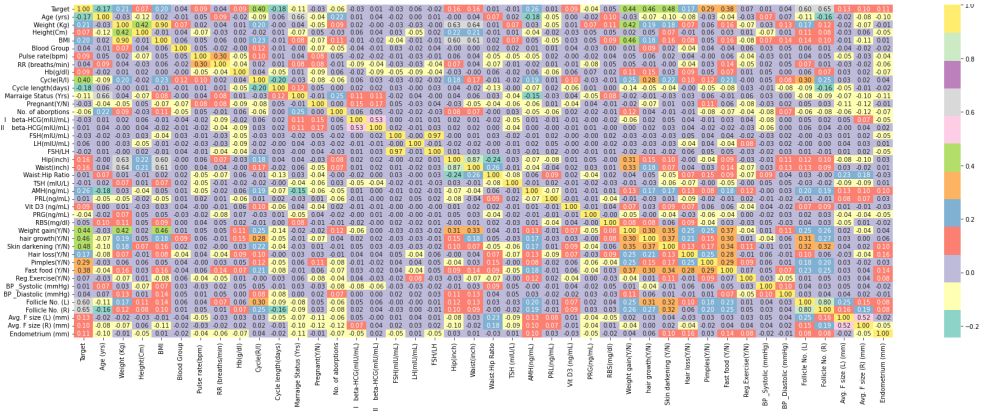


Fig. 14. Correlation Between Features

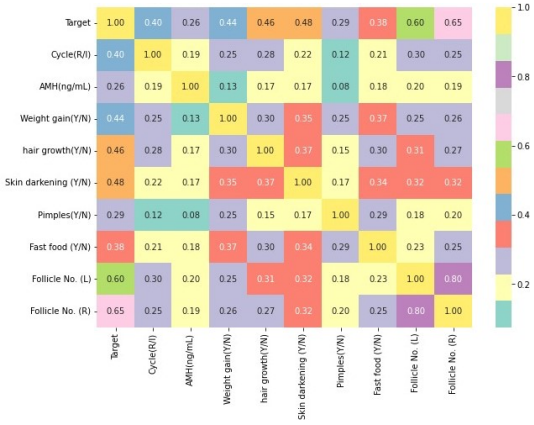


Fig. 15. Correlation Between Features with Corr Theshold 0.25

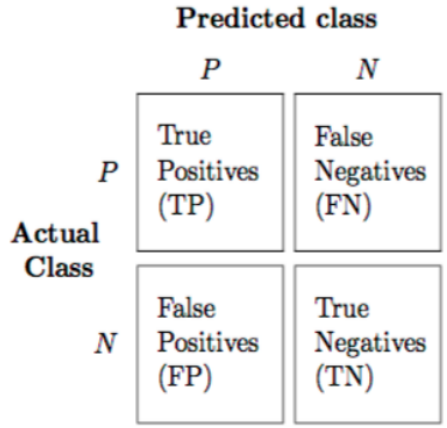


Fig. 16. Confusion Matrix

G. Evaluation of Model

We compared the accuracy of 7 different classification algorithms and compared the mean accuracy of each of them by stratified cross-validation. We implemented machine algorithms such as Decision Tree, SVM, Random Forest, KNN, Logistic Regression, XGBRF, CatBoost.

There are various ways to check the performance of the machine learning models. We have evaluated the model based on the performance measures mentioned below. A confusion matrix is a summary of prediction in a tabular format, that is used to describe the performance of the model. It is a table with a combination of predicted and actual values where,

True Positive (TP) : When a person has PCOS and the model correctly predicts it.

True Negative (TN) : When a person does not have PCOS and the model correctly predicts it.

False Positive (FP) : (Type I Error) When a person does not have PCOS and the model fails to predict it.

False Negative (FN) : (Type II Error) When a person has PCOS and the model fails to predict it.

The following evaluation matrices are as follows:

- Accuracy: It returns the value of how many cases related to PCOS did we correctly label out of all the cases, which is calculated as:

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

- Precision: It returns the value of how many of those who we labeled as having PCOS actually have PCOS, which is calculated as:

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

- Recall or Sensitivity: It returns the value of all the women who have PCOS, how many of those we correctly predict, which is calculated as:

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN}$$

- F1 Score: It is the average of the precision and recall, which is calculated as:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- Specificity: It returns the value of all the women who do not have PCOS, how many of those did we correctly predict, which is calculated as:

$$Specificity = \frac{TN}{TN + FP}$$

- Cross validation accuracy: This approach includes randomly dividing the set of instances into k groups, or folds with equal size. In our work we have take k = 10, 20, 30, 40. and for better practice standard deviation is also calculated for skill score variance.

H. Comparison and selection of the best model

We used seven algorithms and calculated their accuracy, precision, recall, f1 score and specificity. We used these measures to compare the performance pf the model based on the data collected. Higher accuracy of our prediction is considered as a primary metric for comparison whereas higher precision and specificity were also taken into consideration. We realised that if we take 10 features only then good accuracy can be achieved which takes less computation time.

Feature	Score
Target	1
Follicle No. (R)	0.648327
Follicle No. (L)	0.603346
Skin darkening (Y/N)	0.475733
hair growth(Y/N)	0.464667
Weight gain(Y/N)	0.441047
Cycle(R/I)	0.401644
Fast food (Y/N)	0.376183
Pimples(Y/N)	0.286077
AMH(ng/mL)	0.264141
Weight (Kg)	0.211938

Fig. 17. Ranking of 10 best features

V. EXPERIMENTATION AND RESULTS

Total 541 responses were available for this study. The data comprised of women from all age groups. Among the data collected 364 cases were normal non-PCOS cases, while the remaining 177 cases were of women suffering from PCOS.

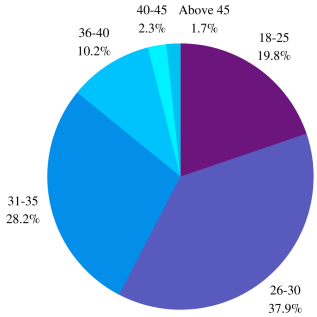


Fig. 18. Age group suffering from PCOS

From this, we can conclude that age group 26-30 are most likely to suffer from PCOS.

The algorithms comprised of 7 models which were Naive Bayes Classifier, K-Nearest Neighbors, XGBRF, CatBoost, Support Vector Machine, Decision Tree Classifier, and Logistic Regression. The below table shows the evaluation metrics used. Based on these metrics, the performance of the models can be analyzed and the best model is selected.

Algorithm	Precision	Sensitivity	F1 Score	Specificity
Decision Tree	0.86	0.86	0.86	0.777777778
SVM	0.613496933	1	0.760456274	0
Random Forest	0.866071429	0.97	0.91509434	0.761904762
KNN	0.803571429	0.9	0.849056604	0.650793651
Logistic Regression	0.652777778	0.94	0.770491803	0.206349206
XBGBRF	0.859813084	0.92	0.888888889	0.761904762
CatBoost	0.886792453	0.94	0.912621359	0.80952381

Fig. 19. Performance of models

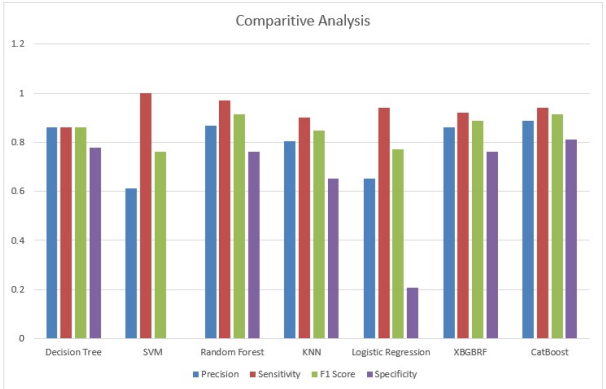
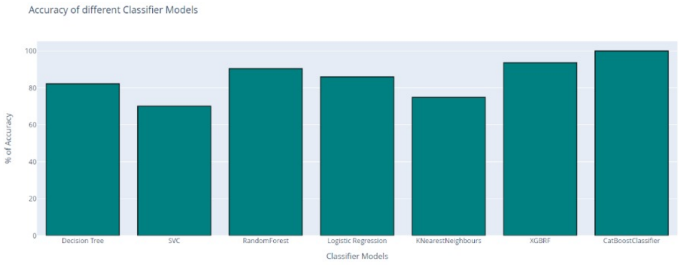


Fig. 20. Comparitive Analysis

From this table and the graph, it's clear that CatBoost performs better than the others.

Now we'll compare the accuracy of different classifier models.



The accuracy of CatBoost is the best according to this graph. Hence, it's clear that **CatBoost** performs the best.

VI. Conclusions and Limitations

Detection PCOS at an early stage enhances the early treatment of the patients. So, this research aims to detect PCOS using Hybrid XGBRF and Catboost models, Gradient Boosting, Random Forest, Logistic Regression, Hybrid Random Forest and Logistic Regression, SVM, Decision Tree depending on various factors. The dataset obtained from Kaggle repository contains 541 patients with 42 attributes. Results showed that attribute Follicle No.(R) is the most important attribute. Results also indicated that if we take 10 features only then good accuracy can be achieved which takes less computation time. The results of all the different classifiers were compared and overall it proved that CatBoost outperformed all the classifiers.

The limitations are that:

A. This research could be validated if we would have a large dataset by having more patients like one thousand. Furthermore, new hybrid algorithms can be produced and if we have a larger dataset then deep learning algorithms like optimized form of CNN can be implemented to increase the classification accuracy.

B. There is a huge scope for this research as cases of PCOS are increasing day by day. We can't use this directly so we can develop a full-fledged application of PCOS Tracker to help the user to keep the track of their symptoms and monitor their activity levels via phone sensors and wearable. The application can also provide data insights and information about PCOS.

References

[1] Preeti Chauhan, Pooja Patil, Neha Rane, Dr. Pooja Raundale, Harshil Kanakia, 2021, June, "Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS". 2021 International Conference on Communication information and Computing Technology

[2] Lele, Priyanka, Thakare, Anuradha., 2020, "Comparative Analysis of Classifiers for Polycystic Ovary Syndrome Detection using Various Statistical Measures". International Journal of Engineering Research

[3] Shakoor Ahmad Bhat, Dr. Rashmi Gupta, 2021, August, "Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms".