Polycystic Ovary Syndrome Prediction [ML]

G.B. Chavan*, Gonte Akanksha**,Hole Swaranjali**,Khande Dnyaneshwari**
Gawade Nikita**

*(Assistant Professor, Information Technology, SVPM's College of Engineering Malegaon(bk),
Baramati Email:ganesh99222@gmail.com)
**(UG Students, Information Technology, SVPM's College of Engineering Malegaon(bk),
Baramati) Email:
gonteakanksha@gmail.com,swaranjalihole20002@gmail.com,dnyaneshwarikhande19@gmail.com
nikitagawade863@gmail.com)

Abstract

Polycystic Ovary Syndrome (PCOS) is a significant health risk for women during their reproductive years. The disorder is characterized mainly by higher levels of male hormones and androgens, which cause the formation of fluid-filled follicles in the ovaries, preventing regular egg release. PCOS can result in difficulties such as miscarriage, infertility, and pregnancy troubles. According to the most recent information, about 31.3% of women in Asia have PCOS, and sadly, 69% to 70% of these instances go misdiagnosed. Recognizing the critical need for research to enable early detection and intervention to prevent serious PCOS effects, our primary research objective is to develop a predictive model for PCOS implementing advanced machine learning techniques. To establish our predictive models, we use a collection of clinical and physical information from women. We provide a novel feature selection method based on an optimized chi-squared (CS-PCOS) mechanism to improve accuracy and efficacy. The Gaussian Naive Bayes (GNB) method emerges as the top-performing model, overcoming other machine learning models and state-of-the-art studies, due to the novel CS-PCOS feature selection technique. GNB provides exceptional results with 100% accuracy, precision, recall, and F1-scores while requiring only 0.002 seconds of calculating time. Our results highlight the importance of various dataset features such as , waist-hip ratio , prolactin (PRL), systolic and diastolic blood pressure, thyroid-stimulating hormone (TSH), relative risk of breaths (RR-breaths), and pregnancy in PCOS prediction. Using the GNB algorithm and these critical criteria, our work intends to aid the medical community in early PCOS detection, thereby reducing miscarriage occurrences and enabling earlier management for women suffering from this disorder.

Keywords — PCOS, women's health, male hormones, infertility, early detection, machine learning, feature selection, Gaussian Naive Bayes (GNB), accuracy, prolactin (PRL), blood pressure, thyroid-stimulating hormone (TSH), pregnancy, miscarriage, intervention

---------------------------------------*************************---------------------------------

I. *I . INTRODUCTION*

Polycystic Ovary Syndrome is the critical disorder in women during their reproductive phase. The hormonal condition known as polycystic ovary syndrome affects the ovaries. For optimal health, the ovaries usually generate estragon, a female hormone, and androgens, a male hormone. Hormones are

substances that regulate bodily activities. The hormones of affected women are out of balance, with either less estragon or more androgens than usual. As a result, fluid-filled sacs known as lumps develops on the ovaries. These lumps become larger over time and eventually get in the way of ovulation. For women with PCOS, this interferes with ovulation, which lowers their chances of getting pregnant. Diabetes, heart disease, high blood pressure, endometrial thickness, sleep apnea, depression, anxiety, eating disorders, and endometrial cancer are just a few of the conditions that women with PCOS are more likely to suffer from. The development of PCOS may also be influenced by environmental variables in addition to genetic ones. The numerous little cysts (fluid-filled sacs) that develop in the ovaries are referred to as polycystic ovarian syndrome. On the other hand, some women without the disease develop cysts, while some women with the disorder do not. A woman may not produce enough of the hormones required for ovulation in certain Situations. In the event that ovulation is unsuccessful, the ovaries may grow several little cysts. Androgens are the hormones produced by these cysts. High levels of testosterone are common in PCOS-afflicted women. This may exacerbate a woman's menstrual cycle issues.

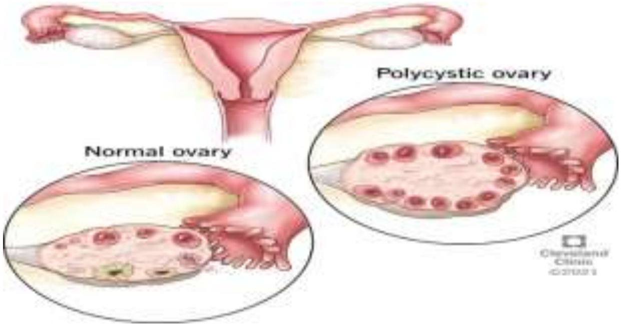

Fig. 1 Difference Between Normal Ovary and PCOS Ovary

Symptoms that could manifest are: periods that are heavy, extended, irregular, unpredictable, or insufficient, acne from infertility or oily skin excessive facial or body hair, weight gain, especially in the abdomen, or baldness with a male pattern, Type 2 diabetes and hypertension are two additional health issues that are more common in people with PCOS. elevated cholesterol heart conditions endometrial cancer, or cancer of the interior of the uterus Male sex hormones called androgens, which are typically present in modest amounts in women, are produced abnormally by the ovaries in patients with polycystic ovarian syndrome (PCOS). The numerous little cysts (fluid-filled sacs) that develop in the ovaries are referred to as polycystic ovarian syndrome. On the other hand, some women without the disease develop cysts, while some women with the disorder do not. PCOS, also known as polycystic ovarian syndrome, is a complicated illness with an unknown origin. Evidence of a genetic link has been found, indicating that it might run in families. Additionally, compared to non-identical twins, identical twins have a higher chance of having PCOS together, according to twin studies. Furthermore, prenatal exposure to elevated hormone levels may raise the chance of having PCOS in later life. It is thought that environmental and genetic variables are involved. The risk can be raised by things like being overweight, not exercising enough, and having a family member with PCOS. In order to make a diagnosis of PCOS, medical professionals need to find at least two of the following three symptoms: high levels of androgen hormones, irregular or non-existent ovulation, and ultrasound-visible ovarian cysts. Hypothyroidism, prolactin excess in the blood, and adrenal hyperplasia are other disorders that present with comparable symptoms. Therefore, when diagnosing a patient, physicians must take these factors into account. Depending on a person's symptoms, objectives, and desire for conception, there are many therapy options for PCOS. It usually entails a combination of medication, lifestyle modifications, and at times surgery. Women with PCOS should collaborate closely with their healthcare professional to develop a customized treatment plan. The disorder can be managed with medication provided by doctors and lifestyle modifications. Doctors are compelled by these diverse symptoms to conduct numerous clinical tests, interpret their results, and recommend unnecessary radiological imaging procedures.

According to popular assumption, machine learning for healthcare is a rapidly emerging field that is becoming more accessible. Many of the concepts behind machine learning require a strong grasp of mathematics and programming. If you understand the principles of machine learning, you can expand these skills to tackle increasingly complicated concepts and issues. This could offer new opportunities for innovation and a range of career paths in the healthcare sector due to machine learning techniques, computers can operate freely without explicit programming. Machine learning applications are fed new data and are capable of self-learning, self-improvement, self-change, and self-adaptation.

II. RELATED WORK

Over the past decade, the detection of Polycystic Ovary Syndrome (PCOS) has emerged as a significant area of research, with various techniques being explored to diagnose PCOS at an early stage. PCOS is a complex endocrine disorder with diverse clinical manifestations, making its diagnosis multifaceted. One crucial diagnostic criterion involves assessing the appearance of the ovaries through ultrasound images, specifically examining factors like the number of follicles, their size, and distribution within the ovaries. Traditionally, this process has required manual tracing and follicle counting on ultrasound images to determine the presence of PCOS. However, a novel method has been introduced to automate the identification of PCOS, streamlining the diagnostic process. This innovative algorithm encompasses several key steps. It begins by segmenting the follicles in ultrasound images, followed by the computation of features based on the segmented follicles using follicle stereology. These features are stored as feature vectors, and the final step involves their classification. The classification involves categorizing the feature vectors into two groups: one indicating the presence of PCOS and the other indicating its absence. This automated tool significantly reduces the time typically consumed by manual tracing and the measurement of each follicle's width and length.

The study employed three different classifiers: the Linear Discriminant Classifier (LDC), k-Nearest Neighbours (KNN), and Support Vector Machine (SVM). Remarkably, LDC outperformed SVM and KNN, although all three classifiers showed promise in their results. This automated approach has the potential to mitigate the risk of serious complications that can arise from delayed PCOS detection. PCOS is characterized by the development of multiple follicular cysts within the ovaries, and the current manual approach of counting these cysts is associated with issues of efficiency, reproducibility, and variability. In response to these challenges, an automated scheme has been proposed for PCOS detection. It initiates by taking an ultrasound image of the ovary as input, which is then processed using an adaptive morphological filter. Next, an adapted labelled watershed model is employed to delineate the contours of the targets within the image. Finally, a clustering method is applied to detect the expected follicular cysts. The evaluation of this automated scheme demonstrated an accuracy of 0.84. However, it's worth noting that due to the specific nature of PCOS, this automated scheme may not be directly applicable to other diverse target identification problems.

In summary, the development of automated methods for PCOS detection represents a significant advancement in the field, offering efficiency and precision that manual methods often lack. These automated approaches have the potential to enhance the early diagnosis of PCOS, reducing the associated risks and complications, ultimately improving the management of this complex medical condition.

TABLE I
SUMMARY DETAILS OF THE RELATED WORK FOR PCOS PREDECTION

Reff No.	Title	Author	Features
1.	Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome	M.M.Hassan	Artificial intelligence can be used in healthcare systems for diag nostic purposes to handle large amounts of clinical data with much accuracy and precision.
2.	Detection of polycystic ovary syndrome using machine learning algorithms	S.A.Bhat	PCOS diagnosis can be tricky, because not everyone with PCOS has polycystic ovaries, nor does everyone with ovarian cysts have PCOS, hence the pelvic ultrasound as a stand-alone diagnosis is not sufficient.
3.	A comparative analysis of logistic regression, random forest and KNN models for The text classification	M.Shah	In the current generation, a huge amount of textual documents are generated and there is an urgent need to organize them in a proper structure so that classification can be performed.

4.	An efficient decision tree establishment and performance analysis with different machine learning approaches on PCOS	A.S.Prapty	Polycystic Ovary Syndrome (PCOS) is an exceedingly serious disease for which a woman has to pay a lot of lifelong damages.

III. METHODOLOGY

Manual PCOS prediction requires a long time. In this research, we're presenting a new automated approach that might be a game changer for early PCOS prediction.

This study focuses on the deployment of the proposed automated technique to help medical competence in PCOS prediction and treatment of PCOS patients. As stated in the literature, the machine Learning methods has proven helpful in Prediction PCOS. A distinct strategy to Prediction PCOS using a machine learning model is proposed in this research. Figure 2 depicts the work flow for research methods, which includes similar actions that will be performed during this research.

As shown in Figure 2, we start by collecting PCOS data. Second, data Pre-processing takes place in which we first cleaning of Database take place. Furthermore, feature selection is used, where important features are selected to predict PCOS in its early stages. The dataset is then split into train and test in a 70:30 ratio. Following that, we construct the model. We will evaluate machine learning methods using results matrices. Then Further we Deploy the model for user interaction. Where User can Input the information and Predict PCOS.

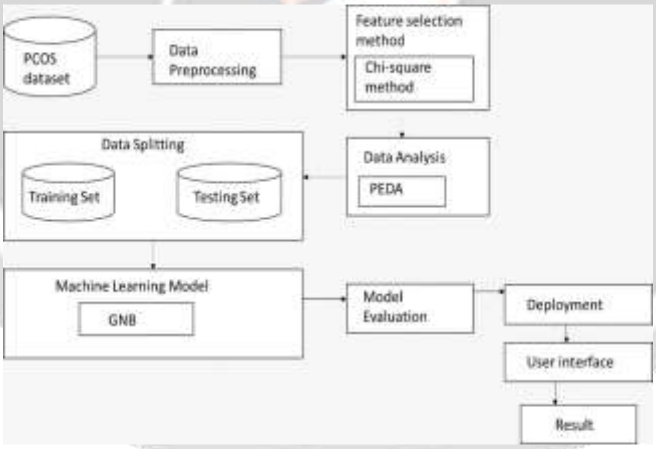

Fig. 2 Research Methodology for Prediction of PCOS

A. ***Data Collection:*** The PCOS Dataset is use to research study in our System. The Clinical and Physical parameter of 541 patients having 41 Features are used to create dataset. This data is collected from 10 different Hospital in Kerala, India. The dataset having import attributes like Waist: Hip Ratio, RR (breaths/min), Cycle length (days), Age (yrs), Weight (Kg), BMI, Marriage Status (Yrs), FSH(mIU/mL), LH(mIU/mL), etc. From all this dataset top 10 highest ranking dataset has to be selected for Early Prediction of PCOS.

B. ***Data Pre-processing:*** This step is take place for Data cleaning and Data Normalization. For Pre-processing NULL values in the dataset are filled with 0. The column which having Unnecessary Information are dropped so the ‘s1_no’ and ‘patient_File_No’ are dropped. This step is important because it takes the machine learning model.

I. ***Data cleaning:*** The method of fixing or eliminating erroneous, damaged, badly formatted, duplicated, or insufficient data from a dataset is known as data cleaning. Even if the conclusions and algorithms appear to be correct, they are useless if the information that underlies them is incorrect. When combining various data sources, there are numerous possibilities for data to be duplicated or incorrectly classified. In general, data cleaning reduces errors and increases the accuracy of data.

II. ***Data Normalization:*** The process of transforming data into the range [0, 1] (or any other range) or simply transforming data onto the unit sphere is referred to as normalization. Normalization gives equal weights/importance to each variable, ensuring that no single variable biases model performance in one direction simply because it is larger. Normalization is essential because it ensures that each feature contributes equally to the training process of the machine learning model. It prevents larger-scale features from dominating the model's decision-making process. We can increase the performance and accuracy of our machine learning models by normalizing the data.

C. ***Feature Selection Method:*** The Feature Selection Method is used to find out or to extract the features. As we all know all features did not have equal importance so we use feature selection model to rank them according to their best fit predictive model with high accuracy. For this purpose we use Chi-Square Method. This Method is used because it can be used to investigate relationship between two categorical variable and determine the importance value of each feature. The CS-PCOS check the importance by comparing the observed frequencies with the expected frequencies.

Further the features which having importance value near to one are all selected as best fit features for PCOS Prediction. The feature which having importance value near to zero are drooped. The features are Age (yrs), Weight (Kg), BMI, Cycle(R/I), Cycle length(days), Marriage Status (Yrs), FSH(mIU/mL), LH(mIU/mL), FSH/LH, AMH(ng/mL), Vit D3 (ng/mL), PRG(ng/mL), Weight gain(Y/N), hair growth(Y/N), Skin darkening (Y/N), Hair loss(Y/N), Pimples(Y/N), Fast food (Y/N), Follicle No. (L), and Follicle No. (R).

I. ***Chi-Square Method:*** Chi square is the statistical method use to find if there is any association between the two categorical variables. For applying the method we must require contingency table, which describes the frequency of distribution of the observed data for any two variables. By using this table we can find the association between two variables. Here how we use this method in our system:

- **Calculate Expected Frequencies(Eij)** for each cell under the assumption that there is no association between the variables.
  $$E_{ij} = \frac{R_i \times C_j}{N}$$
- **Calculate the Chi-square Statistic($\chi 2$)** using the formula
  $$\chi^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
  The summation is over all cells in the contingency table
- **Determine the Degrees of Freedom (df)** using the formula
  df = (R – 1) × (C – 1)

  Where R is the number of rows and C is the number of columns in the contingency table.
- **Find the Critical Value** from a chi-square distribution table or use statistical software, corresponding to the chosen significance level (a) and degrees of freedom.
- **Compare χ^2 and Critical Value to Make a Decision**
  If χ^2 > Critical Value, reject H. There is a significant association between the variables.

  If $\chi^2 \leq$ Critical Value, fail to reject H. There is insufficient evidence to claim a significant association.

D. ***Data Analysis:*** The next step is Data Analysis in which 20 features with top 10 data which rae selected from the earlier are used. We use PCOS Exploratory Data Analysis (PEDA) technique to analyse the data into different angles. EDA is a approach to analyse the data using visual technique and graphical representations. The Seaborn, Pandas, and Matplotlib libraries of python are used to visualize the chart. Various graphs are included to visualize data i.e. histogram, pie chart, 3D scatterplot etc.

E. ***Data Splitting:*** This method include splitting od dataset into train and test dataset. Training dataset is collection of data that is use to train and teach the Machine Learning model or algorithm about how to make prediction or decision. It is compose of input feature and target output. Testing dataset is use to evaluate the model i.e. it is use to determine predicted output and actual output.

F. ***Machine Learning Model Building:*** The next step is model building. Choose an appropriate machine learning algorithm for the task common algorithm for classification task common algorithm for classification tasks including logistic regressions, support vector machine, decision trees, Gaussian Naïve Bayes, Random forests. The Gaussian Naïve Bayes method emerges as the top performance model, overcoming other machine learning models. To train machine learning model we provide them learn pattern and make predictions based on the data they receive during training process.

I. Gaussian Naïve Bayes: If in a data set most of the attributes are continues then Gaussian Naive Bayes is used. It is assumed in this algorithm that predictor values are samples from Gaussian distribution. Hence, Formula for conditional Probability becomes:

$$P(X_i | y) = \frac{1}{\sqrt{2\prod \sigma^2}} \exp\left(- \frac{\quad}{2\sigma^2_y}\right)$$

Here μ_y and σ_y are mean and variance of predictor distribution.

II. ***Naïve Bayes Classifier:*** Naive Bayes is a set of supervised machine learning techniques primarily employed for classification tasks. At its core, it relies on Bayes Theorem to make predictions regarding the likelihood of data points belonging to different classes within a dataset. For each class, the algorithm calculates membership probabilities, estimating the probability that a given data point pertains to a specific class. The class with the highest membership probability is then designated as the most probable class for that data point.
This is expressed as:

$$P(A\backslash B) = \frac{P(B\backslash A)\ P(B)}{P(B)}$$

- C as the class variable representing the class Labels.
- X as the feature vector representing the input variables.

The goal is to find the most likely class C_K given the observed features X.

G. ***Model Evaluation:*** After training, we test i.e. evaluate the performance of machine learning models using various metrics like Accuracy, Precision, Recall and F1-score. This metrics help us access how well the model is perform and make improvement if needed.

I. Accuracy: Accuracy is a union of precision and trueness. Better accuracy means better precision and trueness. It is reported and as uncertainty. Classification Accuracy and Error is calculated as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

II. **Precision:** In classification of PCOS, precision is the fraction of instances allocated to positive class which belong to positive class.

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

III. **Recall:** In classification of PCOS, recall summarizes how accurately the positive class was predicted which means fraction of total amount of related instances which were actually retrieved.

$$\text{Recall} = \frac{TP}{TP + FN}$$

IV. **F1-Score:** F1-Score is the combination of both Precision.

$$\text{F1} - \text{Score} = \frac{2\text{recision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

H. ***Deployment:*** Once the model perform Satisfactorily, deploy it in a destop application to enable PCOS prediction for new patient for early detection of PCOS.

F. ***Result:*** In this section, we discuss complete analysis of machine learning algorithm. After the execution of all the applied models, it is significant to check the performance of each and every model on a training and testing data. The various evaluation metrics have been taken into this research.The performance metrics are the accuracy score, precision score, recall score, and f1-score. The performance metrics are evaluated for scientific validation of our research models. The employed model's accuracy score shows how much the model is good in prediction.

IV.CONCLUSION:

In this research, we explored the use of machine learning algorithms for PCOS PREDICTION. We provided an overview of PCOS, the importance of early prediction, and how machine learning can be used to diagnose the condition. Here we have use Gaussian Naïve Bayes for Prediction Purpose . GNB gives the accuracy of 100%. We hope our research can lead to early detection and more effective treatment of PCOS, which can have a significant impact on Women's health.

ACKNOWLEDGMENT

We take this opportunity to thank our project guide Prof.Chavan.G.B. and Head of the Department Prof.Dr.Gawade J.S. and Honourable Principal Prof.Dr.Mukane S.M. for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report. We are also thankful to all the staff members of the Department of Information Technology of SVPM's College of Engineering, Malegaon(Bk) for their valuable time, support, comments, suggestions *Available at* ***www.ijsred.com*** and persuasion. We would also like to thank the institute for providing the required facilities, Internet access and important books.

REFERENCES

[1] A. Garg and V. Mago, ''Role of machine learning in medical research: A survey,'' Comput. Sci. Rev., vol. 40, May 2021, Art. no. 100370.

[2] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, ''Secure and robust machine learning for healthcare: A survey,'' IEEE Rev. Biomed. Eng.,687 vol. 14, pp. 156–180, 2021.

[3] D. Hu, W. Dong, X. Lu, H. Duan, K. He, and Z. Huang, ''Evidential mace prediction of acute coronary syndrome using electronic health 692 records,'' BMC Med. Informat. Decis. Making, vol. 19, no. 2, pp. 9–17,693 2019.

[4] M. M. Hassan and T. Mirza, ''Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome,'' Int. J. Comput.696 Appl., vol. 175, pp. 42–53, Sep. 2020.

[5] S. Bharati, P. Podder, and M. R. H. Mondal, ''Diagnosis of polycystic ovary syndrome using machine learning algorithms,'' in Proc. IEEE Region 703 Symp., (TENSYMP), Jun. 2020, pp. 1486–1489.

[6] S. A. Bhat, ''Detection of polycystic ovary syndrome using machine learning algorithms,'' M.S. thesis, Dublin, Nat. College Ireland, Dublin,706 Ireland, 2021.

[7] A. Saravanan and S. Sathiamoorthy, ''Detection of polycystic ovarian syndrome: A literature survey,'' Asian J. Eng. Appl. Technol., vol. 7, pp. 46–51,717 Nov. 2018.

[8] V. Thakre, ''PCOcare: PCOS detection and prediction using machine learning algorithms,'' Biosci. Biotechnol. Res. Commun., vol. 13, no. 14,720 pp. 240–244, Dec. 2020.

[9] P. Kottarathil, ‘‘Polycystic ovary syndrome (PCOS) | Kaggle,’’ Hospital,735 Kerala, India, Tech. Rep. 9.71, 2022.

[10] D. T. Barus, R. Elfarizy, F. Masri, and P. H. Gunawan, ‘‘Parallel programming of churn prediction using Gaussian Naïve Bayes,’’ in Proc. 8th Int. 778 Conf. Inf. Commun. Technol. (ICoICT), Jun. 2020, pp. 1–4.

[11] L. Cataldi, L. Tiberi, and G. Costa, ‘‘Estimation of MCS intensity for Italy from high quality accelerometric data, using GMICEs and Gaussian 781 Naïve Bayes classifiers,’’ Bull. Earth. Eng., vol. 19, pp. 2325–2342, 782 Apr. 2021

[12] K. Shah, H. Patel, D. Sanghvi, and M. Shah, ‘‘A comparative analysis of logistic regression, random forest and KNN models for the text classification,’’ Augmented Hum. Res., vol. 5, pp. 1–16, Mar. 2020.