Machine Learning Approaches on Polycystic Ovary Syndrome

[1]Sanjana Mahadik, [2]Vrushali Dhami, [3]Shubhangi Wadibhasme, [4]Rutuja Argade, [5]Prof. Jyotsna Nanajkar
Assistant Professor, Department of Information Technology, Zeal College of Engineering and Research, Pune
UG Student, Department of Information Technology, Zeal College of Engineering and Research, Pune

Abstract - Polycystic ovary syndrome (PCOS) is a common endocrine disorder affecting reproductive-aged women worldwide. It is characterized by a complex interplay of hormonal imbalances, metabolic dysfunction, and ovarian abnormalities. Early detection and diagnosis of PCOS are crucial for timely intervention and management of the condition. This abstract presents an overview of various approaches and advancements in PCOS detection, highlighting both traditional and emerging methods. The traditional diagnostic criteria for PCOS include the Rotterdam criteria, which require the presence of at least two out of three features: irregular menstrual cycles, clinical or biochemical signs of hyper androgenism, and polycystic ovaries observed on ultrasound. However, these criteria have limitations, and newer diagnostic strategies are being explored Keywords: Predictive modeling, Model Training, Diagnosis, Machine Learning.

Index Terms - Polycystic Ovary Syndrome, Support vector machine, Random Forest, Decision Tree and Naive Bayes Classifier.

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a common endocrine disorder affecting by women. It is characterized by hormonal imbalances, enlarged ovaries with multiple cysts, and various symptoms such as irregular menstrual cycles, excess hair growth, and fertility issues [1]. Diagnosing and managing PCOS can be challenging due to its heterogeneous nature and complex etiology. Machine learning techniques have gained popularity in medical research and clinical practice, offering efficient tools for analyzing large datasets and making accurate predictions [2]. Design an efficient decision tree model for the diagnosis and analysis of Polycystic Ovary Syndrome (PCOS). Compare and evaluate the performance of the SVM/decision tree model using different machine learning approaches on PCOS data to identify the most accurate and reliable classification technique for PCOS diagnosis [3] [4]. PCOS is often under-diagnosed or diagnosed late, leading to delayed treatment and potential health complications. Developing an accurate and reliable method for early detection can help identify PCOS at its early stages, allowing for timely intervention and management [5]. PCOS detection project aims to improve the quality of life for women with PCOS by enabling early diagnosis, personalized treatment, and comprehensive support. It has the potential to positively impact reproductive health, mental well-being, and long-term outcomes for individuals affected by this syndrome.

II. RELATED WORK

Polycystic ovary syndrome is common most of the women in the world like African and white women (5%-8%), Spain(6.8%-13%) [3] and it is most common in Asian country [4].

A woman affected by PCOS has some common syndrome like hair fall, unusual blood pressure, excessive weight, menstrual cycle length, pregnancy [3]. For detecting Polycystic 3D ultra-sounds and Pulsed-Doppler ultrasounds have an efficient result [16]. Automated screening for detecting the presence of PCOS has been done with Logistic Regression and Bayesian classifier [14]. But our result is more satisfying. Another study for PCOSidentification was done using an ultrasonography, clinical, andendocrine factor based on feature Follicle No, Period cycle length primary infertility but the number of data is only 60 patients which is too low to go for valuable decision [13].A morphological image processing filter with a watershed algorithm was applied in the ultrasound image to detect the PCOS [17]. The main purpose of

our study to find the presenceof PCOS in a patient applying different machine learning methods like SVM, Decision Tree and Random Forest Algorithm for better accuracy than the performed Random Forest and Naive-Bayes classification.

III. DATA COLLECTION

For applying our machine learning approaches a dataset is driven from ten individual hospitals where data contains a patient's physical health condition to regulate PCOS related problems. We have a total of 542 data where 177 people are affected by PCOS and the rest of the patients are normal. Clinical features are the main focus on the feature mostlyrelatable for PCOS [8]. The menstrual cycle plays a vital role to define PCOS as an irregular and infrequent cycle of the periodwhich is the first symptom of PCOS. The dataset contains 31 individual features for the confirmation of having PCOS. Have BMI 25 is considered overweight according the World Health Organization (WHO) [23]. Overweight is another prominentissue for PCOS diseases. Hair growth, hair loss, weight gain are some more visible changes [9]. Other dominant features are respiratory rate, hemoglobin rate in blood, menstrual cycle length, pregnancy, and marital status, follicle-stimulating hormone, luteinizing hormone, thyroid hormone, anti-mullein hormone level, prolactin level, and endometrium and RBStest. Imbalanced hormones are needed to test. Throughout our research, we guarantee to establish the most significant features that are required to keep an eye on for every womanto prevent at an early stage [10].

IV. SYSTEM ARCHITECTURE

Preprocessing part a balanced dataset is made by reducing the abundant class size. Preprocessing is pursued by dividing the data into a training set and testing set for model creation. We handle 70% of data for training purposes and the rest for testing the accuracy of the selective models. Fig. 1 describes our proposed system architecture. The following figure high-lights 2 parts, one is the implementation of different methods of machine learning for performance evolution and the other is establishing the top most responsible features for PCOS.

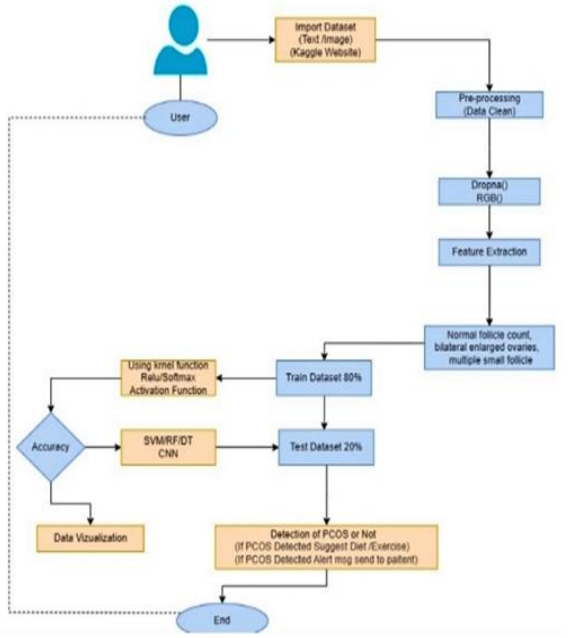

Fig.1. System Architecture

V. IMPLEMENTATION OF DIFFERENT CLASSIFIERS

SVM, Naive classifier, and Random Forest methods are implemented and finally, a comparison among them is produced deciding the best performer. The prediction is about whether a patient is affected by PCOS or not.

A. SVM

Support vector machine (SVM) is the most powerful ma- chine learning algorithmic model used as a binary classi- fication for minimizing generalization error by construction decision boundary explicitly [5] [6]. As SVM is more suitable forthe complex dataset it performs better than any other machine learning model. The classification of SVM is linearly separableand the decision boundary is adjusted by a hyper plane between the nearest points of each categorical class. Regarding the other methods, it has a target factor whether a patient suffers from PCOS or not that is numerically presented as 1 and 0 respectively. The rest of the columns (follicle-stimulating hormone, luteinizing hormone, thyroid hormone, anti-mullein hormone level, prolactin level, etc.) are acted as predictors. Having this target value and predictors the dataset has split 70% into the training set and 30% into test set. For gettinga better accuracy level the normalization is done to keep the data range between [0-1].

The normalization formula is below: Normalize Value = (Value - Train min) / (Train max - Train min). The

accuracy is generated by SVM if best for normalize factor and accuracy is 87%. SVMs can be used in PCOS diagnosis, especially when dealing with small datasets, feature-based analysis, linearly separable data, or a need for interpretable results. Two-way classifier we use Linear Kernel [17].

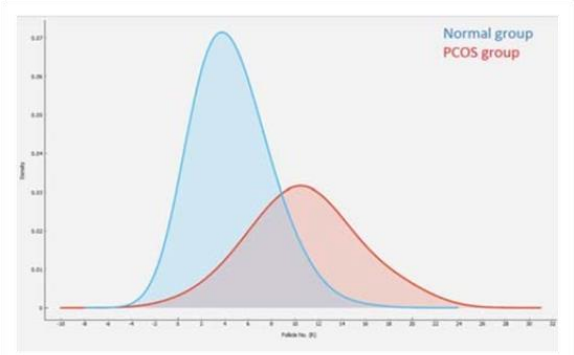

Fig. 2. Left Follicle No

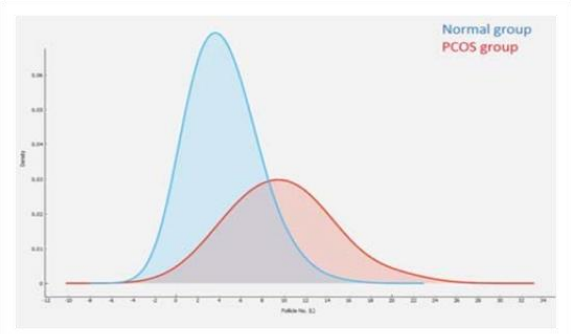

Fig. 3. Right Follicle No

B. NAVIE CLASSSIFIER

It is a technique of classification based on Bayes' Theorem, assuming individuality among predictors [12]. In a simple sense, the existence of a specific feature in a class is not related tothe presence of another feature in a Naive Bayes classification. Implementing this classifier to our PCOS dataset, this gives a good result.

C. RANDOM FOREST

It creates a bootstrapped sample from the original dataset and creates a lot of decision trees in which the subset ofvariables is used randomly at each step. As with bootstrap sampling, several rows can be picked several times and many of them never, most of the time such out-of-bag samples are tested and always correctly labeled. It can be strongly mentioned that the Random forest classifier plays best and ends up with a solid conclusion as it considers every created decision tree's opinion based on multi voting [11]. To define regression problems through the random forest, the Random Forest Repressor class of the sklearn ensemble library is used. Parametern estimators sets the number of trees in the random forest. Theresults of our algorithm are observed from n estimator=100 and finally ended with n estimator=10000 where it works perfectly.

D. COMPARISON

Implementing these four methods, which method performs best can be compared. For that accuracy, precision, recall and f1-score are considered as the parameters. The comparison among them is shown in Table [14]. SVM, Naive classifier, and Random Forest are implemented and SVM gavethe best result. In this paper, some other methods are also implemented.

TABLE -PERFORMANCE COMPARISON AMONG 4 METHODS

Method	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1 score (%)*
SVM	*87*	*86*	*82*	*84*
Random Forest	*81*	*83*	*78*	*80*
Decision Tree	*76*	*75*	*74*	*75*
Naive Bayes	*93*	*80*	*80*	*80*

But Random forests can be concluded with the best of these classifiers based on this overall performance. Thus, further analysis was done with the help of the Random Forest. First, adecision tree is developed for each sample using it and select the best result from the prediction of those decision trees. Then the topmost responsible features are finally established. These most valuable attributes need to be approached with cautionto avoid PCOS at the early stage.

E. NAIVE BAYES

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset.

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

The formula for Bayes' theorem is given as:

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

F. DECISION TREE

It could be used when interpretability, feature importance, and visual representation are essential. They can provide a starting point for building more complex models or be used in ensemble methods to improve accuracy and generalization.

They can provide a starting point for building more complex models or be used in ensemble methods to improve accuracy and generalization. One-way classifier we use Entropy (messy data are split based on values of the feature vector associated with each data point.

Using the Random Forest method, a decision tree is developed. Shannon Entropy and Information Gain are the keys. The tree is shown in Fig. 4 in two parts.

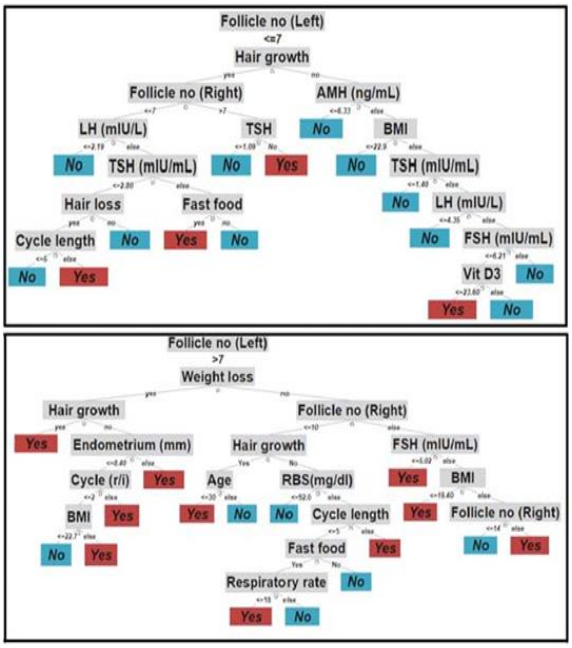

Fig. 4 Decision Tree

G. VALIDATION OF MODELS

The common performance evaluation metrics for validation of models include:

Accuracy: - It is the proportion of the total number of predictions that were correct and can be calculated from the following equation:

Accuracy =Tx+Ty Tx+Fx+Ty+Fy

Where, Tx= True Positives, Fx= False Positives, Ty= True Negatives, Fy= False Negatives

Recall: - It is defined as the percentage of total relevant results correctly classified by the algorithm.

Recall=Tx/Tx+Fx

Precision: - refers to the percentage of the results which are relevant.

Precision=Tx/Tx+Ty

F-statistics:-It is a metric that combines precision and recall and is calculated as the harmonic mean of precision and recall.

Fn=(1+n^2) *precision*recall/((n^2precision) +recall)

H. FEATURE EXTRACTION

Feature extraction is a crucial step in PCOS detection, as it involves selecting and transforming relevant information from the raw data that can be used as input to machine learning or deep learning models.

In the context of PCOS detection, you can extract features from various data sources, including clinical data and medical images (such as ultrasound scans).

Here's a detailed explanation of feature extraction for PCOS detection:

1. Cycle Length: Calculate the length of menstrual cycles.

2. Cycle Regularity: Assess the regularity of menstrual cycles using metrics like standard deviation of cycle length.

3. Hormone Levels: Extract hormone levels, including FSH (Follicle-Stimulating Hormone).

4. Glucose and Insulin: Extract fasting glucose and insulin levels.

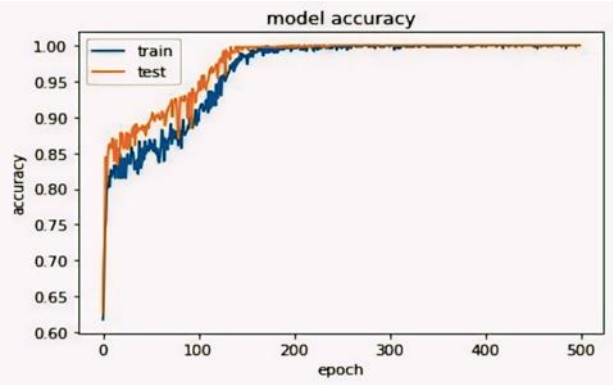

Fig. 5 Model Accuracy

VI. EXPERIMENT RESULT

PCOS detection project is the ability to identify the syndrome at an early stage. Early detection allows for timely intervention and management, which can help prevent or minimize the development of complications associated with PCOS. Early intervention can also lead to more effective treatment outcomes and improved quality of life for individuals with PCOS. Personalized treatment, improved reproductive health outcomes,

long-term health management, and contributions to scientific research. By detecting PCOS early and implementing appropriate interventions, individuals with PCOS can experience better health outcomes and quality of life. PCOS shares symptoms with other hormonal disorders and health conditions, such as thyroid disorders or adrenal gland abnormalities. These overlapping symptoms can complicate the detection process and lead to misdiagnosis or delayed diagnosis of PCOS. Distinguishing PCOS from other conditions with similar symptoms requires careful evaluation and exclusion of alternative causes. PCOS detection projects can help identify PCOS at an early stage, allowing for timely intervention and management. Early diagnosis enables healthcare providers to initiate appropriate treatments, lifestyle modifications, and counseling to address the symptoms and prevent long-term complications associated with PCOS. By detecting PCOS early, individuals can receive the necessary support and treatment to improve their quality of life

VII. CONCLUSION

PCOS detection projects contribute to education and awareness, both among healthcare providers and the general public. They enhance understanding of PCOS symptoms, diagnostic criteria, and management strategies, leading to improved access to healthcare services and support for affected individuals. By focusing on early detection, personalized treatment, fertility management, long-term health, education, and research, these projects can contribute to better healthcare outcomes and increased awareness of PCOS, ultimately leading to improved quality of life for those affected by this syndrome. Future work in the field of Polycystic Ovary Syndrome (PCOS) detection is expected to bring about significant advancements in early diagnosis, treatment, and overall management of this prevalent endocrine disorder. Collection and curation of large and diverse datasets to improve the robustness and generalizability of PCOS detection models. This would involve collaboration with healthcare institutions and data sharing initiatives.

REFERENCES

[1] P. Mehrotra, C. Chakraborty, B. Ghoshdastidar, S. Ghoshdastidar and K. Ghoshdastidar, "Automated ovarian follicle recognition for Polycystic Ovary Syndrome," 2011 International Conference on Image Information Processing, Shimla, India, 2019, pp. 1-4, doi: 10.1109/ICIIP.2011.6108968.

[2] S. S. Deshpande and A. Wakankar, "Automated detection of Polycystic Ovarian Syndrome using follicle recognition," 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram, India, 2020, pp. 1341-1346, doi: 10.1109/ICACCCT.2014.7019318.

[3] P. Soni and S. Vashisht, "Exploration on Polycystic Ovarian Syndrome and Data Mining Techniques," 2019 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 816-820, doi: 10.1109/CESYS.2018.8724087.

[4] Adiwijaya et al, “Follicle Detection on the USG Images to Support Determination of Polycystic Ovary Syndrome”, Journal of Physics Conference Series 622(1):012027

[5] Agrawal A., Ambab, R. Lahoti, R. Muley, P & Pande. P. S.(2022).Roles of artificial intelligence in PCOS detection . Journal of Data Megha Institute of Medical Sciences University .17(2),491.

[6] Kodipalli A., & wisesty, U.N. (2018, March). Classification of polycystic ovary based on ultrasound images using competitive neural network. In journal of Physics: Conference Series (Vol.971,No. 1, p.012005).IOP Publishing.

[7] Brattain L. J., Telfer B. A., Dhyani, M., Grajo j. R., &Samir A. E. (2018).ML for medical ultrasound: Status, methods and future opportunities. Abdominal radiology ,43(4),786-799.

[8] Pulluparambhi. S. J., & Bhat. S. (2021). Medical Image Processing: Detection and Prediction of PCOS –A Systematic Literature Review. International Journal of Health Sciences and Pharmacy (IJHSP),5(2),80-98.

[9] Gautam N. Allahbadia, Rubina Merchant, "Polycystic ovary syndrome and impact on health," Middle East Fertility Society Journal, vol.16. pp. 19-37, October 2010.

[10] S. franks, "Medical progress: polycystic ovary syndrome," New England Journal of Medicine, vol.333(13):853-861, 1995.

[11] Asa Lindholm, Liselott Andersson. Mats Eliason. Marie Bixo, Inger Sundstom-Poromaa. “Prevalence of Symptoms associated with Polycystic

Ovary Syndrome." International Journal of Gynaecology and Obstetrics, vol. 102, pp.39-43, January 2008.

[12] Lin Li, Dongzi Yang, Xiaoli Chen, Yaxiao Chen, Shuving Feng, Liangan Wang. "Clinical and Metabolic feature of Polycystic Ovary Syndrome." International Journal of Gynaecology and Obstetrics, vol.97, pp.129-134, January 2007.

[13] 1.F. Stein, M. Leventhal, "Amenorrhea associated with bilateral poly- cystic ovaries," American Journal Obstetrics and Gynaecology, 1935; 29: 181-191.

Ricardo Azziz. Keslie S. woods, Rosario Reyna. Timothy J. Key Eric S. Knochenhauer, Bulent O. Yildir, The Prevalence of the Polycystic Ovary Syndrome in an Unselected Population." The Journal of Clinical Endocrinology and Metabolism, vol.89(6), pp.2745-2749, 2004,

[14] Ovalle F, Azziz R. "Insulin Resistance, polycystic ovary syndrome and Type 2 diabetes mellitus," Fertility and Sterility, vol.77, pp1095-1105, 2002.

[15] Wild RA. "Long term health consequence of PCOS," Human Reproduction Updates, vol.24, pp231-241, 2002. Legro RS." Polycystic Ovary Syndrome and Cardiovascular Disease:

[16] premature association? "Endocrine Reviews, vol24, pp.302-312, 2003. P Hardiman, OS Pillay, W Atiomo "Polycystic Ovary Syndrome and Endometrial Carcinoma," Lancet, vol.361, pp. 1810-1812. 2003.

[17] Maryruth J. Lawrence, Mark G. Eramian, Roger A. Pierson, Eric Neufeld, "Computer Assisted Detection of Polycystic Ovary Morphology in Ultrasound Images," IEEE Transl. Computer Society, 2007[Fourth Canadian Conference on Computer and Robot Vision, 2007.

[18] Gerard S Convay," Polycystic Ovary Syndrome: Clinical Aspects," Bailer's Clinical Endocrinology and Metabolism, vol. 10, 1996.

[19] sWafaa M Aboul Enien, Nadia A Barghash, Fayrouz S Mohamed Ali. Clinical, "Ultrasonography and Endocrine predictors of Ovarian response to Clomiphene Citrate in Norm Gonadotropic an ovulatory Infertility." Middle East Fertility Society Journal, vol.9. 2004.

[20] Palak Mehrotra Jyotirmoy Chatterjee, Chandan Chakraborty Biswanath

[21] Ghoshdastidar, Sudarshan Ghoshdastidar," Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques" Annual IEEE India Conference, 2011.

[22] Morva Tahmasbi Rad." BMI role in treatment of infertile patients with polycystic ovary syndrome." International Congress Series, vol.1271. pp.34-37,2004.

[23] A. H. B elen, J.S.E. Laven S-L Tan, D. Dewailly. "Ultrasound Assessment of the Polycystic Ovary: international Consensus Definition's," Human Reproduction Updates, vol. 6, pp.505-514,