PCOS Detection using Machine Learning Algorithms

Diya Gandhi
Northeastern University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Bansri Patel
Pace University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Namrata Dave
G H Patel College of Engineering and Technology (GCET)
13 PUBLICATIONS **495** CITATIONS
SEE PROFILE

RESEARCH ARTICLE

PCOS Detection using Machine Learning Algorithms

Diya Gandhi[1*], Bansri Patel[1] and Namrata Dave[1]

[1]GH Patel College of Engineering and Technology, CVM University, V.V Nagar, Anand, Gujarat, India.

Received: 30 Dec 2023 Revised: 09 Jan 2024 Accepted: 12 Jan 2024

***Address for Correspondence**
Diya Gandhi
GH Patel College of Engineering and Technology,
CVM University,
V.V Nagar, Anand, Gujarat, India.

ABSTRACT

Polycystic Ovary Syndrome (PCOS), is a hormonaldis order that occurs among women in their reproductive age. Ithas effective conflicts throughout this gynecological disorder, as it affects one inten women at a nearly age. There are certain symptoms such as irregular menstrual cycles, missed periods, heavy bleeding during the menstruation period, excess of and rogen hormones, obesity, acne or oily skin, hair growth on the face, and a typical weight gain. The exact cause of PCO Sis not yet properly defined, but it could involve genetic causes and anim balance in the diet. Due to certain effectiveness like the risk of heart attack, and type two diabetes, it is necessary to get detected and diagnosed as early as possible and start the possible treatments which include a healthy diet and exercises, with medications like birth control pills that control the level of hormones. Certain Machine Learning algorithms are used to detect this disorder. The data set consists of 541 patients, and out of 44 features, 10 potential features were identified using the filter method. This paper includes a detection model of PCOS using various machine learning algorithms like Random Forest, Logistic Regression, Support Vector Classifier, and Decision Tree. Among all these algorithms, Random Forest has 83.48% accuracy for the model.

Keywords: Polycystic Ovary Syndrome, Machine Learning, Random Forest, Logistic Regression, Support Vector Classifier, Decision Tree.

INTRODUCTION

Technology is boosting its measure every single time which makes every transformation very flexible whether it is in the gadgetsor the health care industry and services. Machine Learning plays a paramount role in all health-related domains as it is a constituent subset of artificial intelligence. There are distinct application areas it such as image recognition, health monitoring, robotic perception, anomaly detection, and many more. It predominantly focuses on the development of algorithms that can be easily accessible from the data sets that are provided for detecting and predicting the required information. Thus, Machine Learning algorithms are utilized efficiently for the detection of PCOS. PCOS is a common hormonal disorder observed in women of child bearing age. Few symptoms indicate the hormonalim balance, and it results in obesity associated with an enlarged polycystic ovary. In there productive age of 15-40, women experience their regular trend of their menstruation with hormonal effects, which shows that PCOS can affect individuals at any age. There are certain health risks due to this disorder including cardiovascular diseases which generally increase blood pressure and cholesterol levels, end ometrial cancer occurs because the least ovulation leads to the build up of the uterinelining, mental health issues affect physiological conditions such as depression and anxiety, and type two diabetes happens due to insulin resistance and high blood sugar levels increase the risks of circumstances. The significant element in this heterogenous condition is hormones. Luteinizing Hormone (LH), Follicles-Stimulating Hormone (FSH), and Anti-Mullerian Hormone (AMH) affect ollicles and the development of the eggs, creating issues in ovulation, and FSH levels might be normal or lower than the usual values. Estrogens and Progesterone are essential for balancing the level of hormones to get the regular menstrual cycle. Among every suffering patient, 70% are undiagnosed Hence, the prediction and detection of PCOS is necessary at the preliminary phase as it sustains the life of an individual by reducing lifelong health risks and creating a healthy life style.

The certain work focused in this paper is:

I. Selection of the influential components affecting thepatients of PCOS with the help of feature selection.
II. Implementation of various machine learning algorithms on the selected features of the dataset. Comparing the accuracy of the different algorithms tofit the best model

LITERATURE REVIEW

PCOS detection has become a hot topic for researchers in the last decade. Few individuals have implemented the various methodologies in this field to achieve the desired outcome for the health benefit to all women

This section consists of the distinct literature works done previously based on various implemented methods such as follicles detection, feature extraction, and classification, Cross Validation, Support Vector Machine (SVM), Logistic Regression, k nearest neighbors (kNN), and many more [4].

METHODOLOGY

Data Collection

Data collection is a crucial step. For this, various platforms areavailable example for Kaggle, UCI Repository etc. In thispaper, we have used a dataset from Kaggle [1]. This dataset is composed of 44 different features with more than 500 records.Such features include pimples, hair growth, cycles, vitamin d3,etc.

Data Preprocessing

Data Preprocessing is a step that takes raw data and transformsit into a format that can be understood and analyzed. Unprocessed data must contain some Missing values, Outliers,Unstructured manner, and Categorical data. Missing values canbe corrected in many ways but the most common methods are Delete Rows with Missing Values and replace the missing value with some arbitrary value using fillna(), Missing values can also be imputed using ‘interpolation’. Here we have also dropped unnecessary features. Furthermore, the dataset should only contain a value that is float or integer so that algorithms can process the data. The next step is Exploratory data analysis. This process involves summarizing, visualizing, and getting deeply acquainted with the important traits of a dataset. It examines a correlation matrix of all the features, and how all the features correlate with the PCOS, having a look at features bearing significant correlation.[4].

Feature Selection

The feature selection method intends to select the most useful feature for a model to predict the output. Feature selection is performed to improve predictivity, reduce the dimensionality of feature space, and get rid of noisy data.

Some favored techniques for feature selection are Filter Methods, Wrapper Methods, and Embedded methods. In this paper, we have used the filter method to rank each feature based on some univariate metric and then select the highest-ranking features and we havealso referred to previous research to select the highest-ranking features [5].

Fitting into models

After the Data preprocessing, it is now ready to be handled by the models. Selected sets of features are used to study the algorithm. Among countless ML algorithms available, we haveapplied Logistic Regression (LR), Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Support Vector Machine.

Logistic Regression (LR)

Logistic Regression, a supervised learning algorithm, uncoversits preliminary application in classification tasks by assessing the probability of a sample belonging to a distinct class. It is specifically fitted for binary classification, where the output variable is categorical. This algorithm operates the logisticfunction, also known as the sigmoid function, to convert the result of a linear equation into a value within the range of 0 to 1. This altered value represents the likelihood or probability ofa data point being associated with a certain class. [11]. The accuracy of this algorithm was 82.56% here

Decision Tree Classifier

The Decision Tree classifier is a supervised algorithm principally operated for classification tasks. This technique operates by iteratively splitting the dataset into subsets according to the attribute values, resulting in a tree-like configuration. In this structure, individual inner node exemplifies a conclusion based on a distinctive characteristic, and each leaf node corresponds to a class label. Here, the accuracy of this algorithm was 77.98%.[12]

Gradient Boosting Classifier

Gradient Boosting is a significant boosting approach that assembles numerous weak learners into vital learners. This methodology involves training individually unique samples to minimize the loss function, such as mean squared error or cross-entropy, based on the performance of the previous model employing gradient descent. In each iteration, the algorithm computes the gradient of the loss function regarding the predictions assembled by the current ensemble. Thereafter, a unique weak representative is trained to minimize this gradient. The predictions yielded by the new model are incorporated into the ensemble, and this iterative approach persists until a predefined stopping criterion is satisfied. The accuracy here was 82.56% [12].

Random Forest Classifier

The Random Forest Algorithm is a supervised machine learning technique employed for addressing both classification and regression challenges in the realm of machine learning. It can be considered as an ensemble of decision trees. Instead of depending on a single decision tree, the random forest contains multiple decision trees, each prepared on distinct subsets of the delivered dataset. To enhance predictive accuracy, the algorithm computes the intermediate prediction from these trees. Instead of just depending on one tree's outcome, the absolute prediction is determined by a majority vote among the predictions from the ensemble of trees. The accuracy for this algorithm was 83.48%.[13]

Support Vector Machine

The Support Vector Machine (SVM) is a supervised learning algorithm appropriate for both classification and regression tasks, although it is primarily employed in classification problems in the field of machine learning. The primary objective of SVM is to establish an optimal conclusion limitation, usually directed to as a hyperplane, within an n- dimensional distance to effectively distinguish between different classes. This hyperplane relieves the proper categorization of further data attributes in the future. SVM identifies the critical data points that play a major role in determining this hyperplane; these pivotal representatives are known as support vectors, giving rise to the name "Support Vector Machine”. The accuracy was 70% here.[14]

Evaluation and Comparison of Models

The comparison of these models is done based on accuracy. Various classification algorithms are used to find the most acceptable models. As shown in the table and plot the best accuracy is given by Random Forest Classifier, Gradient Boosting classifier, and Logistic Regression

RESULT

The dataset contained 541 samples with 44 features. Out of these 44 parameters, only ten parameters are considered. Parameters that are more important for the diagnosis of PCOS are shown in Table III, after analyzing the performance of all five models, we can conclude Random Forest is most Suitable.

CONCLUSION

This paper exhibits the different Machine Learning algorithmsand a model to detect the early phase of PCOS, as it is essential for women's health. This hormonal disorder impacts the regular condition of women and disturbs the psychological, physical,and metabolic components. Day-to-day exercise and a regular healthy diet are initialized to decrease the effect and maintain a nourishing lifestyle. The model in this paper ventures thecomfortable system to detect the disorder at an early stage, with a definitive set of parameters. Among all the various algorithmsused, the Random Forest Classifier possesses the foremostresult in its performance with 83.48% by considering the relevant 10 features. This model is flexible such that it can be utilized by doctors for the early detection of PCOS. Hence, wehave built the model with different machine-learning techniques to detect PCOS at an early stage

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who supported and assisted us with their determination and invaluable contribution that led to the completion of our paper. We are grateful to the medical professionals who shared their insights and expertise by giving us their valuable time. We are also thankful to our college as they supported us generously with their essential resources.

REFERENCES

1. https://www.kaggle.com/datasets/ayamoheddine/pcos-dataset
2. Sandy Rihana, Hares Moussallem, Chiraz Skaf, and Charles Yaacoub. Automated algorithm for ovarian cysts detection in ultrasonogram. In 2013 2nd International Conference on Advances in Biomedical Engineering, pages 219222, 2013.doi:10.1109/ICABME.2013.6648887
3. Bedy Purnama, Untari Novia Wisesti, Adiwijaya, Fhira Nhita, Andini Gayatri, Titik Mutiah, "A Classification of Polycystic Ovary Syndrome Based on FollicleDetec-tion ofUltrasound Images, 2015 3rd International Conference on Information and Communication Tech-nology (ICoICT).
4. Amsy Denny, Anita Raj, Ashi Ashok, Maneesh Ram C, Remya George, "i-HOPE: Detection and Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques", 2019 IEEE Region 10 Conference (TENCON 2019).
5. Subrato Bharati, Prajoy Podder, M. Rubaiyat Hossain Mondal, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms". 2020 IEEE
6. Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh.
7. Madhumitha, J., M Kalaiyarasi and Sathya Ram. "Automated Polycystic Ovarian Syndrome Identification with Follicle Recognition." 2021 3rd InternationalConference on Signal Processing and Communication(ICPSC) *(2021): 98-102.*
8. Pijush Dutta, Shobhandeb Paul, Madhurima Majumder et al. An Efficient SMOTE Based Machine Learning

classification for Prediction &amp; Detection of PCOS, 08 November 2021, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs- 1043852/v1]

9. Tiwari, Shamik, Lalit Kane, Deepika Koundal, Anurag Jain, Adi Alhudhaif, Kemal Polat, Atef Zaguia, Fayadh Alenezi, and Sara A. Althubiti. "SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning." *Expert Systems with Applications* 203 (2022):117592.
10. Samia Ahmed; Md. Sazzadur Rahman; Ismate Jahan; M. Shamim Kaiser, A. S. M Sanwar Hosen, Deepak Ghimir, Seong-Heum Kim "A Review on the Polycystic Ovary Syndrome using machine learning" in *IEEE Access*, vol. 11, pp. 86522- 86543,2023, doi: 10.1109/ACCESS.2023.3304536.
11. Hosmer, D. W., Lemeshow, S., & Sturdivant, R.
12. X. (2013). Applied Logistic Regression. John Wiley &Sons.
13. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. CRC press.
14. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189-1232.
15. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
16. Cortes, C., & Vapnik, V. (1995). Support- vector networks. Machine learning, 20(3), 273- 297.

Table 1. Research Methodology

AUTHORS	OBJECTIVES	RESEARCH METHODOLOGY	RESULTS
Rihana etal. [2013] [2]	classificationn in ovary and Cysts detection, ultrasound images with geometricalfeatures of the cyst.	Image pre-processing,Feature extraction, SVM classifier, andValidation were used byROC.	Accuracy of 90% was achieved and cysts were detected inovary ultrasound images.
Purnama et al. [2015][3]	Detecting follicles via ultrasound (USG) pictures through a process involving binary follicle images, feature extraction, and segmentationn.	Multiple classification methods were developed suchas SVM – RBF kernel, Neural Network – LVQ, and KNN – Euclideandistance.	At K=5,KNN attained an accuracy of 78%, and on C=40, 82% accuracy was achieved in the SVM-RBF kernel.
Denny et al. [2019][4]	Diagnosis of PCOS based on dataset available on Kaggle.	Attributes of PCOS are transformed with PCA by various machine learning algorithms such as Decision Trees, Random Forest, SVM, KNN, etc.	Random Forest was the best model for PCOS detection with an accuracyof 89%.
Subrato et al. [2020][5]	Diagnosis of PCOS using Kaggle dataset.	Algorithm used for classification are gradient boosting, Random Forest , Logistic regression, RFLR and used holdout and cross validation methods	RFLR gave highest accuracy of 91.01% with 90% recall value
Madhumt ha et al. [2021][6]	Used image segmentation to get details of the ovary for example follicle size, type of cysts.	SVM, KNN and Logistic Regression were used as per pre- processing and morphological operations.	With the combination of all three algorithm, the hybrid model gave 0.98 accuracy.
Pijush et al. [2021] [7]	Detection and prevention of PCOS.	The algorithm used were SMOTE and five other algorithms Logistic Regression, Random Forest, Support vector machine and K- NN, and Random Forest together for early detection of PCOS.	The best model achieved, Recall: 98%, Precision: 98% and AUROC: 95.6%.

Shamik Tiwari et al. [2022] [8]	To diagnose PCOS using Machine Learning	The algorithms used for classification are SVM, DT, RF, LR, GB, AB, XB, AND CB for correlation coefficients of various levels.	Random Forest (RF) gave highest accuracy of 93.25%
Samia Ahmed et al. [2023] [9]	A review on the PCOS using the Machine Learning	A study on various dataset used for PCOS diagnosis was conducted. In quantitative and Qualitative approaches, the performance of algorithms are compared.	The shortcomings like insufficient dataset, lack of clustering approach, not were detected in this paper.

Table 2. Accuracy of all Models

Models	Accuracy
Logistic Regression	82.56%
Decision Tree Classifier	77.98%
Gradient BoostingClassifier	82.56%
Support Vector Machine	70%
Random Forest Classifier	83.48%

Table 3. Selected Features

Ranking	Features name	Value
1	FSH/LH	Between 1 and 2 (normal),2 or 3(abnormal)
2	FSH (mIU/mL)	4-8(abnormal)
3	AMH (ng/mL)	1-4 (normal), >4 (abnormal)
4	BMI	<24 (normal), >24 (abnormal)
5	Weight gain (Y/N)	Yes(y)/No(n)
6	Follicle No. (L)	<12(normal) >=12(abnormal)
7	Follicle No. (R)	20-30(abnormal)
8	Avg. F size (L) (mm)	2–9 mm in diameter
9	Cycle	(Regular/Irregular)
10	Cycle Length	Number of days

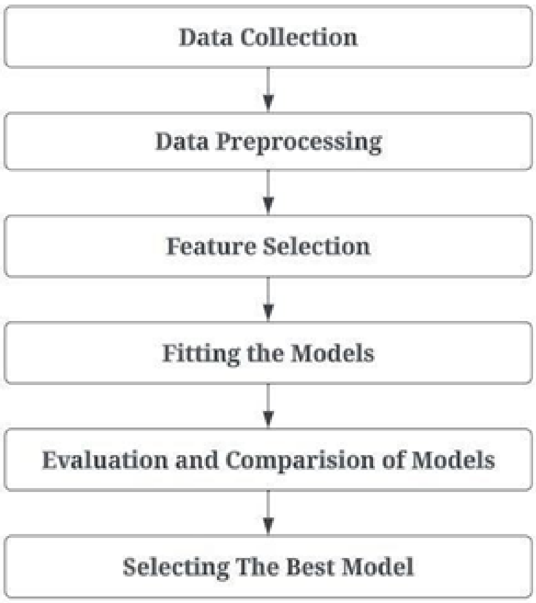

Fig I: System Flow of the Model

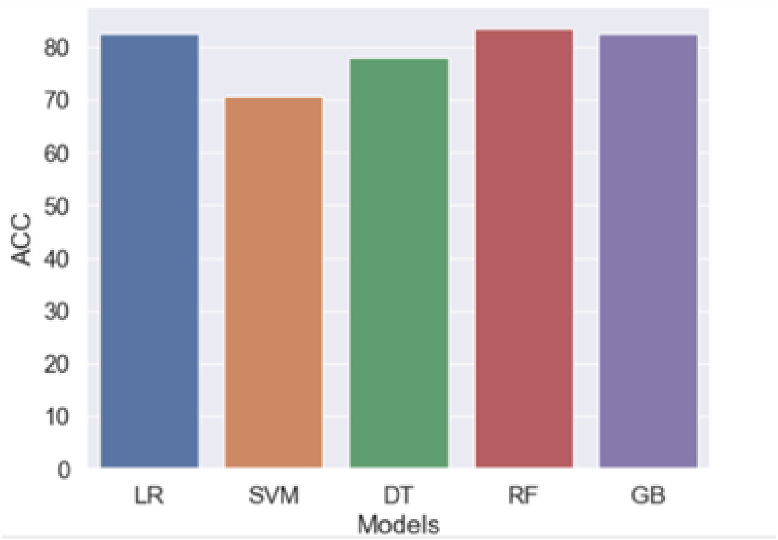

Fig 2.: Accuracy of all Models