A Review on PCOS Diagnosis Using Data Mining Algorithms

Ardra Nair[1, a)], Geetika Chatley[2,b)] and Komal Arora[2, c)]

[1]*Research Scholar, Department of Computer Science, Lovely Professional University, Punjab, India*
[2]*Assistant Professor, Department of Computer Science, Lovely Professional University, Punjab, India*
[a)] *ardra333nair@gmail.com*
[b)]Geetika.25107@lpu.co.in
[c)]*komal.17783@lpu.co.in*

Abstract. This review paper explores how data mining algorithms might help with PCOS diagnosis, a condition that affects 8 to 20% of females worldwide who are of propagative age. Even though PCOS is a complex endocrinopathy that can have negative effects, it is frequently misdiagnosed. Because of the condition's symptoms, such as irregular menstrual cycles, infertility, hirsutism, acne, and obesity, women may experience a poorer quality of life. A questionnaire that identifies women who need PCOS diagnostic tests may help with an early diagnosis and identification of the disease because there is currently no known cure for PCOS. However, due to the heterogeneity of the syndrome and the absence of specific biomarkers, diagnosing PCOS can be difficult. Data mining technologies have been used more frequently recently in the healthcare industry to extract usable information from sizable datasets and speed up the decision-making process. Data mining can be quite effective at predicting PCOS using a brief questionnaire by utilizing machine learning techniques. The focus of this review is on exploring several typical data mining techniques that have the potential to be utilized in diagnosing PCOS.

Keywords: Polycystic ovary syndrome, Data Mining, Decision Support System, Classification, Clustering, Regression

INTRODUCTION

Polycystic ovary syndrome (PCOS) is a hormonal illness that disturbs females of childbearing age. It is distinguished by ovarian cysts, irregular menstruation cycles, and excessive levels of male hormones (androgens). PCOS can cause a wide range of symptoms and health issues, such as infertility, weight gain, acne, excessive hair growth, and insulin resistance. The production of androgens by the ovaries is abnormally high in PCOS. This leads to an imbalance in the hormones that control reproduction. Small cysts (fluid-filled sacs) may develop on the ovaries as a result of an absence of ovulation. Nevertheless, despite the label "polycystic," PCOS may not necessarily involve ovarian cysts. The most common indicators of PCOS are menstrual irregularities, such as amenorrhea or infrequent periods. It could also result in significant bleeding during periods. Excessive hair growth: Examples of abnormal hair growth include hirsutism and thick hair development on the hands, chest, and belly. This may influence about 70% of PCOS-afflicted females [23]. Even as adults, acne can still exist and is notoriously problematic to treat. Additionally, almost 80% of females with PCOS are overweight making it difficult for them to lose weight. Discoloration of Skin: Particularly in the wrinkles of the neck, armpits, and groin, then under the breasts, dark skin patches are seen. Cysts: Small fluid pockets arecommon in PCOS patients' ovaries of fluid are common in the ovaries of PCOS patients. Skin tags: Skin tags are

little, protruding skin flaps. They are frequently found on the neck or in the armpits in women with PCOS. Hair loss: People with PCOS may experience patches of hair loss or begin to grow bald.

Female infertility is most frequently caused by PCOS. Lack of ovulation or decreased ovulation frequency can prevent conception. PCOS can exist without any apparent symptoms. PCOS frequently stays undetected until a person experiences problems with infertility or unexplained weight gain. Another possibility is mild PCOS, which might not have any obvious symptoms. There is no recognized cause for PCOS. There is proof that genetics are involved. It has been demonstrated that having PCOS increases the chance of developing several illnesses, including diabetes, elevated blood pressure in the cardiovascular system, hyperplasia of the endometrium, carcinoma of the uterus, and problems of sleep, like sleep apnea, and both anxiety and depression [23]. Finding women who are at risk for PCOS can be aidedby asking simple self-screening questions like terminal hair growth with a male pattern, depilatory practices, and obesity. This has important implications for assisting women and medical professionals in appropriately identifying those who need additional testing for this endocrinopathy. When predicting PCOS, a number of data mining techniquesare available to forecast the relationship between various symptoms.

LITERATURE SURVEY

The authors developed a 4-item questionnaire with a multivariate logistic regression specificity of 94% and a sensitivity of 77% to detect PCOS. In addition to internal validation using a bootstrap approach, a second sample of 117 occurrences was used to validate the expected accuracy. The authors have created a simple-to-use clinical tool to help in PCOS diagnosis. [5]A 4-item questionnaire designed by the authors to identify PCOS has a multivariate logistic regression specificity of 94% and a sensitivity of 77%. The expected accuracy was validated using a second sample of 117 occurrences in addition to internal validation using the bootstrap approach. An easy-to-use clinical tool has been developed by the authors to aid in PCOS diagnosis.[5] In order to diagnose polycystic ovarian syndrome (PCOS), the author [8] compared multiple machine-learning techniques. The decision tree method was shown to be the most useful in the study for diagnosing PCOS. The study sheds light on how machine learning methods might be used to diagnose and treat PCOS, which could ultimately lead to better patient outcomes.[11] In the study [14], it is discussed how PCOS can be automatically detected using machine learning approaches. To increase the precision of PCOS detection, the authors suggest a classification model utilizing support vector machines (SVMs) and feature selection methods. The work emphasizes the opportunity for early and precise PCOS diagnosis with machine learning. In the work in [15], probabilities for the diagnosis of PCOS were predicted using an ensemble of six models, including SVM, logistic regression and ridge classifiers, bagging classifier, gradient boosting classifier, and random forest classifier. The models' outputs and the starting data were pooled, and a neural network was trained on the combined dataset for final classification. The proposed model achieved an accuracy of 90.74%. The research describes an innovative PCOS detection system that makes use of machine learning methods. The suggested approach seeks to increase PCOS diagnosis precision and lessen reliance on arbitrary clinical judgments. The model had a 93% accuracy rate.[15]The paper[16] presents a new approach for predicting Polycystic Ovary Syndrome (PCOS) using machine learning techniques in bioinformatics. The proposed system utilizes features extracted from gene expression data of PCOS patients and healthy individuals to train a machine-learning model. The model achieved an accuracy of 97.47% in identifying PCOS cases, outperforming traditional diagnostic methods.

The paper[17] presents a comparative analysis of different machine-learning algorithms for the prediction of PCOS. The study utilized clinical and demographic data of PCOS patients and healthy individuals to train and evaluate various algorithms. The results showed that the Random Forest algorithm outperformed other algorithms in terms of accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve. The article[18] explores using data to identify PCOS in women using several machine learning classifiers. The KNN classifier had the maximum sensitivity, and the Linear Discriminant classifier had the highest accuracy, according to

the authors' diagnostic model constructed using MATLAB. The authors propose a classification model that uses various algorithms, to classify patients as either having PCOS or not. The study[18] evaluates the performance of these algorithms using a dataset of 400 female patients and finds that decision trees perform the best .The study [19] used a dataset of 541 women, 177 of whom have PCOS, and finds that FSH/LH level the utmost vital feature. The researchers test different classifiers and approaches, comparing how well they predict PCOS. In the study[20], 100 women had samples of their follicular fluid taken, and the researchers employed effective feature selection methods and Raman spectroscopy to develop an automated model that separates PCOS-positive and PCOS-negative patients. The outcomes demonstrate that this method is 100% accurate at predicting PCOS.

TABLE1.Comprehensive Review of Prior Research

Sr. No	Author	Year	Algorithm	Findings
1	Sinthia, G., Poovizhi, T., & Khilar, R.	2022	Linear Regression, K-means,SVM	Support Vector Machine (SVM) algorithm was the most effective with an accuracy of 91%
2	Nasim, S., Hussain, M. A., Riaz, F., Iqbal, N., & Qazi, A.	2022	Gaussian naive bayes	The GNB algorithm attained 100% accuracy and a minimum computation time of 0.002 seconds
3	Hdaib,D et al	2022	SVM,Neural Network, Naïve Bayes, Classification Tree, Logistic Regression, Linear Discriminant	The most effective classifier, in terms of accuracy, precision, and specificity, was the linear discriminant classifier. But the KNN classifier delivered the greatest results when measuring sensitivity.
4	Reka, S. and Elakkiya,R.	2022	Raman Spectroscopy with advanced ML algorithms	Model using Raman spectra and advanced ML algorithms achieved 100% accuracy using follicular fluid samples.
5	Neto C et al.	2021	Logistic Regression, Multilayer Perceptron, Neural Networks, Random Forest, and GNB	The best model, which made use of RF, produced acceptable results, with an accuracy of 0.95
6	Katarya, R., Srivastava, G., & Chauhan, N.	2021	Ensemble model	The proposed system achieved 90.74% accuracy
7	Chauhan, P., Rani, M., Jain, N., & Jain, N.	2021	KNN, Naive Bayes, Decision Tree, SVM, and LR	The model found to be utmost accurate was the decision tree.
8	Thomas, N., & Kavitha, A.	2020	KNN, SVM, linear regression.	SVM had the peak accuracy at 91%, while KNN was at 75%, K-means at 72%, and linear regression at85%.
9	Bharati, S et al.	2020	Hybrid random forest and logistic regression (RFLR), Gradient boosting, random forest, logistic regression	Through the utilization of 40-fold cross-validation to divide the data into training and testing portions, RFLR was found to have the highest accuracy of 91.01%
10	Mehrotra, P et al.	2011	Bayesian classifier, Logistic regression	Bayesian classifier outperformed logistic regression with an accuracy of 93.93%

DATA MINING TASKS

The central theme of data mining involves extracting essential information from data sets using various data mining techniques such as prediction and classification. Two main types of data mining tasks are descriptive (defining general qualities of data) and predictive (using inference on current data to forecast future behavior). Examples of data mining tasks include classification, prediction, correlation analysis, association, clustering, summarization, and outlier analysis. These tasks help uncover meaningful insights from raw data and can be functional in several fields such as fraud detection and customer behavior analysis. By applying data mining techniques to large and complex datasets, it becomes possible to extract insights that would be challenging or impossible to uncover using traditional methods. Data mining tasks can help organizations make informed decisions, develop better strategies, and gain a competitive advantage in their respective fields.

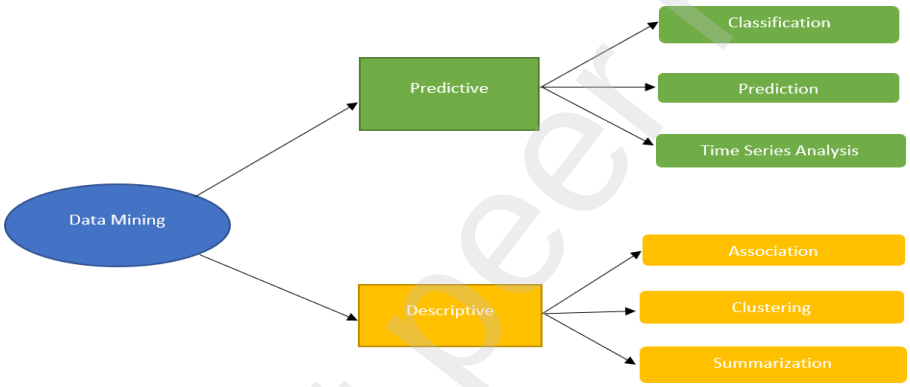

FIGURE1.Data Mining Task

CONCLUSION

Over half of the 10 million people with PCOS are unaware of their condition, indicating a need for increased public awareness of PCOS [21]. By increasing awareness, individuals may better recognize symptoms and seek necessary medical attention. Although PCOS is difficult to cure, certain therapies can help alleviate specific symptoms. A questionnaire could be used to identify those who require diagnostic tests for PCOS, potentially reducing the number of undiagnosed and untreated cases. Data mining is an upcoming breakthrough in medicine, offering the opportunity to discover patterns in healthcare data that could aid in patient diagnosis. A standardized questionnaire survey combined with predictive data mining algorithms could be effective in predicting PCOS and improving diagnostic efficiency. In later studies, the effectiveness of using a patient's basic clinical data and lifestyle to predict outcomes rather than relying on more complex medical assessments can be explored. This method might be used as a first indicator by doctors to decide whether additional testing is required.

REFERENCES

1. Bedrick, B. S., Eskew, A. M., Chavarro, J. E., & Jungheim, E. S. (2020). Self-administered questionnaire to screen for polycystic ovarian syndrome. Women's Health Reports, 1(1), 566-573.
2. Subha, R., et al. (2022). Computerized Diagnosis of Polycystic Ovary Syndrome Using Machine Learning and Swarm Intelligence Techniques.
3. Thomas, N., & Kavitha, A. (2020). Prediction of polycystic ovarian syndrome with clinical dataset using a novel hybrid data mining classification technique. International Journal of Advanced Research in Engineering and Technology, 11(11), 1872-1881.
4. Barber, T. M., McCarthy, M. I., Wass, J. A. H., & Franks, S. (2006). Obesity and polycystic ovary syndrome. Clinical endocrinology, 65(2), 137-145.
5. Pedersen, S. D., Brar, S., Faris, P., & Corenblum, B. (2007). Polycystic ovary syndrome: validated questionnairefor use in diagnosis. Canadian Family Physician, 53(6), 1041-1047
6. Soni, P., & Vashisht, S. (2018, October). Exploration on polycystic ovarian syndrome and data mining techniques. In 2018 3rd International Conference on Communication and Electronics Systems (ICCES) (pp. 816-820). IEEE.
7. Cheng, J. J., & Mahalingaiah, S. (2019). Data mining polycystic ovary morphology in electronic medical recordultrasound reports. Fertility Research and Practice, 5(1), 1-7.
8. Sinthia, G., Poovizhi, T., & Khilar, R. (2022). Analysis on Polycystic Ovarian Syndrome and Comparative Study of Different Machine Learning Algorithms. In Advances in Intelligent Computing and Communication: Proceedings of ICAC 2021 (pp. 191-196). Springer Nature Singapore.
9. Vijayalakshmi, N., & UmaMaheshwari, M. (2016). Data mining to elicit predominant factors causing infertilityin women. Interantional J. Comput. Sci. Mob. Comput, 5(8), 5-9
10. Munjal, A., Khandia, R., & Gautam, B. (2020). A machine learning approach for selection of polycystic ovariansyndrome (PCOS) attributes and comparing different classifier performance with the help of weka and pycaret.Int J Sci Res, 9, 1-5.
11. Neto, C., Silva, M., Fernandes, M., Ferreira, D., & Machado, J. (2021). Prediction models for Polycystic OvarySyndrome using data mining. In Advances in Digital Science: ICADS 2021 (pp. 210-221). Springer InternationalPublishing.
12. Chatley, G., Kaur, S., & Sohal, B. (2016). Software clone detection: A review. International Journal of Control Theory and Applications, 9(41), 555-563.
13. Bulsara, J., Patel, P., Soni, A., & Acharya, S. (2021). A review: Brief insight into Polycystic Ovarian syndrome.Endocrine and Metabolic Science, *3*, p.100085.
14. Mehrotra, P., Chatterjee, J., Chakraborty, C., Ghoshdastidar, B., & Ghoshdastidar, S. (2011). Automated screening of polycystic ovary syndrome using machine learning techniques. In 2011 Annual IEEE India Conference (pp. 1-5). IEEE.
15. Katarya, R., Srivastava, G., & Chauhan, N. (2021). A Novel Polycystic Ovarian Syndrome Diagnostic System Using Machine Learning. In Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020 (pp. 333-343). Springer Singapore.
16. Nasim, S., Hussain, M. A., Riaz, F., Iqbal, N., & Qazi, A. (2022). A novel approach for polycystic ovary syndrome prediction using machine learning in bioinformatics. IEEE Access, 10, 97610-97624.
17. Chauhan, P., Rani, M., Jain, N., & Jain, N. (2021). Comparative analysis of machine learning algorithms for prediction of PCOS. In 2021 International Conference on Communication Information and Computing Technology (ICCICT) (pp. 1-6). IEEE.
18. Hdaib, D., Alkafaween, E., Alzoubi, K., & Khalayleh, W. (2022). Detection of Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms. In 2022 5th International Conference on Engineering Technologyand its Applications (IICETA) (pp. 1-6). IEEE.
19. Bharati, S., Podder, P. and Mondal, M.R.H., 2020. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In: 2020 IEEE Region 10 Symposium (TENSYMP). IEEE, pp. 1486-1489.

20. Reka, S. and Elakkiya, R., 2022. Early Diagnosis of Poly Cystic Ovary Syndrome (PCOS) in young women: A Machine Learning Approach. In: 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). IEEE, pp. 286-288.
21. https://www.unicef.org/india/stories/do-pcod-and-pcos-mean-same-thing-or-are-they-different
22. https://www.derby.ac.uk/blog/why-more-awareness-is-needed-for-polycystic-ovary-syndrome/http://www.austinfitmagazine.com/November-2014/What-isPolycycstic-Ovary-Syndrome-PCOS/
23. https://my.clevelandclinic.org/health/diseases/8316-polycystic-ovary-syndrome-pcos