iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.57876

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

A Systematic Review on: Polycystic Ovary Syndrome Using Machine Learning

Sanika Satpute[1], Vaishnavi Patil[2], Amruta Kutte[3], Janhavi Sonar[4], Mrs. Swati Rajput[5]
Department of Computer Engineering, Dr. D. Y. Patil Institute of Engineering, Management and Research, Maharashtra, India

Abstract: ***Polycystic Ovary Syndrome (PCOS) is a medical condition which causes hormonal disorder in women in their childbearing years. The hormonal imbalance leads to a delayed or even absent menstrual cycle. Women with PCOS majorly suffer from excessive weight gain, facial hair growth, acne, hair loss, skin darkening and irregular periods leading to infertility in rare cases. Early detection and diagnosis of PCOS are crucial for timely intervention and management. The existing methodologies and treatments are insufficient for early-stage detection and prediction. Machine learning can play a significant role in automating PCOS detection based on medical data and patient information. To deal with this problem, we propose a system which can help in early detection and prediction of PCOS treatment from an optimal and minimal set of parameters. In this study, we present an online application that enables women to conveniently estimate their likelihood of having the condition while remaining in the comfort of their own homes until assistance becomes accessible. Using feature selection approaches, we select only the non-invasive, readily measurable characteristics at home, resulting in a minimal but optimal set of clinical data for the same prediction. We employ the random-forest classifier, which has demonstrated encouraging outcomes when used in conjunction with machine learning techniques to diagnose PCOS. The Kaggle dataset, which consists of data points with clinical variables such as weight increase, LHS levels, hair loss, acne, BMI, and follicles in the left and right ovary, is used in this work.***
Keywords: ***PCOS, Random Forest, Decision Tree, Machine Learning, Diagnosis***

I. INTRODUCTION

A prevalent endocrine condition affecting people of reproductive age, mostly women, is called polycystic ovary syndrome, or PCOS.

Hormonal abnormalities, irregular menstrual periods, and many tiny cysts on the ovaries are its defining features. Numerous health concerns, such as infertility, diabetes, cardiovascular disease, and mental health problems, can result from PCOS. The main symptoms of PCOS in women include excessive weight gain, facial hair development, acne, hair loss, skin discoloration, and, in rare instances, irregular periods that can result in infertility. It has been demonstrated that altering one's lifestyle effectively manages PCOS. To completely avoid the possibility of any future health problems, PCOS management is essential.

The etiology of PCOS is still unknown because it is a very complicated condition. PCOS can be caused by genetic, behavioral, lifestyle, or environmental factors. It is frequently characterized by inflammation and insulin resistance. Approximately 80% of individuals with PCOS experience irregular cycles. High testosterone is seen in about 7 out of 10 PCOS patients. Excessive testosterone levels can result in hirsutism, or face and body hair, acne, and hair loss on the head. Small ovarian cysts are a symptom of PCOS rather than the cause of the condition, thus not everyone with PCOS gets them. PCOS occurs often. It may impact around 1 in 12 women and individuals with reproductive age cycles (8%, or 6–13%), however this probably varies according on the community.

A rising number of people are interested in using artificial intelligence and machine learning techniques for medical diagnostics. Machine learning algorithms are able to handle and analyze vast amounts of patient data, spot intricate patterns, and deliver more precise and unbiased outcomes. Machine learning has the potential to enhance the precision and efficacy of PCOS diagnosis, hence facilitating prompt intervention and improving patient outcomes. This method includes analyzing many data sources, including clinical and hormonal factors, ultrasound findings, and patient demographics, using machine learning algorithms. These data points are used to train algorithms that determine if a person has PCOS or not.

Machine learning algorithms look at these variables and how they relate to each other in an effort to find patterns that point to PCOS. The goal is to develop a tool that will help medical professionals diagnose patients more quickly and accurately. The use of such a tool has the potential to mitigate the subjectivity that is inherent in conventional diagnostic techniques and offer a more uniform approach to PCOS diagnosis.

II. LITERATURE REVIEW

Machine learning for diagnosis of PCOS This paper authored by Dr R. Rekha, Ms. Srinithi V. Introduces among the LR, KNN, and RFR, the SMOTE based Random Forest algorithm had the highest accuracy rate of 99.10 percent. The authors suggested that datasets with PCOS and non-PCOS categories are highly imbalanced, and the SMOTE technique was utilized to compare accuracy. The overfitting issue caused by random oversampling is reduced with the aid of SMOTE technique. The CNN model is trained by using image datasets to more effectively classify disorder.

Another Paper, authors Dr. Pooja Raundale, Harshil Kanakia, Pooja Patil, Neha Rane, Preeti Chauhan - Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS. Their research provides different machine learning techniques, such as KNN, Naïve Bayes, SVM, LR and RF are used to classify PCOS. As per the accuracy and confusion matrix, the Decision Tree Classifier emerged as the most accurate model for prediction PCOS. The Decision Tree Classifier demonstrated the highest performance, achieving an accuracy of 81%, precision of 70%, and specificity of 94%. The Gini Importance, which is used to provide scores to input characteristics to a predictive model, was used to examine the best features to determine their relevance.

A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics. This paper is authored by Shazia Nasim, Mubarak Almutairi, Kashif Munir, Ali Raza, And Faizan Younas. This paper employs PEDA (PCOS Exploratory Data Analysis) for feature engineering, incorporating the CS PCOS feature selection approach and correlation analysis. A unique 3D analysis delves into the distribution of features by class, providing valuable insights. The study concludes with a comparative performance evaluation of machine learning models, utilizing accuracy, precision, recall, and F1-score as key metrics. This approach offers a thorough assessment of the models' efficacy in addressing PCOS, contributing to informed decision-making for early detection and effective management.

Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Technique. This paper is authored by Yasmine A. Abu Adla, Dalia G. Raydan, Mohammad-Zafer Charaf, Roua A. Saad, Jad Nasreddine, Mohammad O. Diab. This paper consists of a streamlined yet comprehensive approach to data preprocessing, which involves the essential tasks of detecting and classifying outliers and errors. Employing feature selection techniques such as ANOVA and Sequential Forward Floating Selection enhances the dataset's relevance by isolating key features for subsequent modeling. The study further employs the K-Fold Cross Validation technique for training and evaluating the model, ensuring robust performance assessment. This succinct yet thorough methodology underscores the paper's commitment to delivering reliable insights for effective data analysis and modeling.

PCOS Perception analysis prediction using Machine learning algorithms. This paper is authored by Sakthipriya Dhinakaran, Chandrakumar Thangavel, Shivayavashilaxmipriya S, Harinee V S. This paper methodology aids in the prediction of early symptoms and physical changes in women's bodies. This has led to the development of feature selection, a variety of machine learning algorithms, including logistic regression and K-nearest neighbor (KNN), decision trees that employ decision-making principles, and additional SVM, Naive Bayes, and other approaches like Random Forests for the gathered data set. The models were created once the top six attributes—out of the twenty-one—were chosen. The outcomes demonstrate that these six features can be used to achieve good accuracy.

Another Paper, Accessible Polycystic Ovarian Syndrome Diagnosis Using Machine Learning, which is authored by Akanksha Tanwar, Anima Jain, Anamika Chauhan. The goal of this project is to simplify the diagnosis of polycystic ovarian syndrome for all women. To make this goal attainable, the five most crucial and readily quantifiable characteristics for PCOS prediction are chosen. Additionally, the study's Random Forest Classifier detects PCOS with an accuracy of 92.59%.

III. METHODOLOGY

A. Data Pre-processing

One of the most important steps in getting raw data ready for machine learning models is data preparation. It entails doing things like encoding categorical variables, scaling features, and handling missing values. The quality of information provided is improved by proper data preparation, which produces more accurate and dependable model results.

1) *Data Cleaning:* The goal of data cleaning is to find and fix mistakes or inconsistencies in datasets. Taking care of outliers, eliminating duplicate entries, and fixing formatting problems with data are typical duties. A clean dataset is one that is acceptable for analysis or model training, accurate, and consistent.
2) *Feature Selection:* Feature selection involves choosing the most important features of the model, while discarding unimportant or unnecessary ones. This helps improve model performance, reduce overfitting, and improve interpretability. Techniques include filtering methods (eg, correlation analysis), wrapper methods (eg, recursive feature elimination), and embedded methods (eg. LASSO regression).

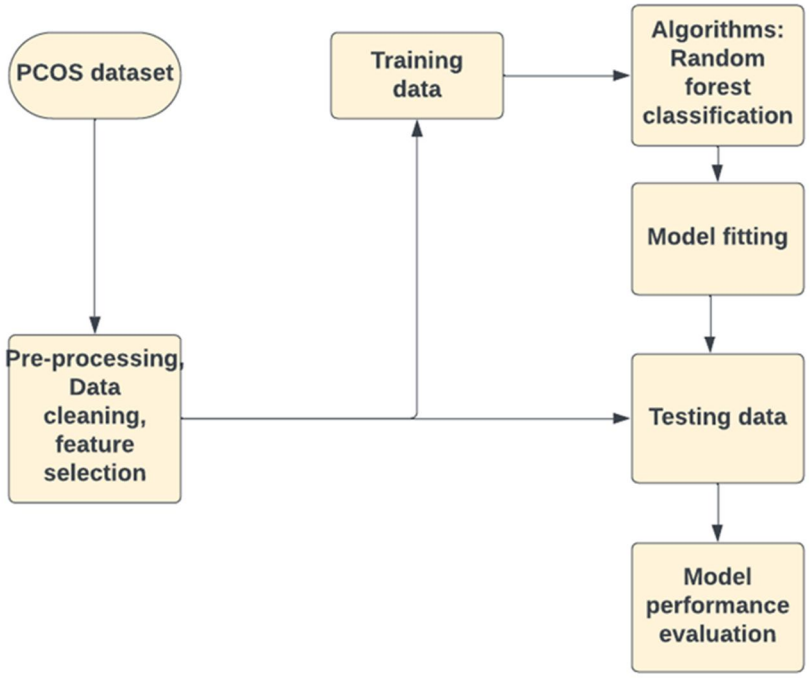


Fig 3.1 PCOS Assistant using Machine Learning Flowchart

B. Machine Learning Techniques

1) *Random Forest classifier:* The Random Forest (RF) is a robust supervised classification model that assembles a forest of multiple decision trees. These trees are constructed randomly, leveraging data samples to form decision nodes representing features and tree leaf nodes signifying the target output. The final prediction is determined through a majority voting process, where the aggregated decisions of individual trees contribute to the ultimate outcome. Crucially, the Gini index and entropy play pivotal roles in guiding the data splitting process within the nodes of the trees, ensuring effective and informed partitioning.
2) *SVM:* One type of supervised machine learning model is the support vector machine (SVM). Regression and classification issues are the principal uses for the SVM. Finding the optimal decision boundary in an n-dimensional feature space to divide data points into the appropriate categories is the main goal of the SVM model. The hyperplane is the optimal decision boundary in support vector machines (SVM). In order to create the hyperplane, the SVM model chooses the extreme vectors.
3) *Decision Tree:* A decision tree is a kind of supervised learning system that uses graphical representation to identify potential solutions to an issue based on specified parameters. Decision points are represented by nodes in the tree structure, and possible outcomes are led by branches. Classification and Regression Tree, or CART, is the algorithm that Decision Trees most frequently use.
4) *Logistic Regression:* For binary classification tasks, the supervised machine learning model Logistic Regression (LOR) is utilized. Using training data that includes independent factors, this model predicts the categorical dependent variable. Most importantly, the target class needs to be represented as a discrete value. The LOR model produces probabilistic values as outputs, ranging from 1 to 0, in contrast to linear regression (LIR).
5) *KNN:* A simple, non-parametric machine learning model, the k-Nearest Neighbors (KNN) approach is mostly applied to classification issues. New input points are assigned to a category similar to those adjacent in the KNN model based on the computation of similarity between data points. Based on how closely new data resembles data in the training set, the model predicts new data based on stored data. Interestingly, KNN is regarded as a lazy learner model since it waits until classification time to learn from the training set. In contrast to enthusiastic learners who pick up a model during training, KNN generates predictions instantly.

C. Train /test Split

We initially divided the data into two primary sets, the training set and the testing test, in order to implement all of the techniques mentioned above. We train our model on a training dataset and validate its prediction by testing the model that was generated on a testing dataset. Using the Python library, we added the split to the models we constructed.

D. Model Performance Evaluation

Model performance evaluation is a critical step in machine learning, as it evaluates the effectiveness of a trained model in predicting new, unseen data. This process compares the model and predictions with actual results, typically using metrics such as accuracy, precision, recall, and F1 scores. In addition, evaluation may include methods such as cross-validation to ensure reliability and generalizability. A deep understanding of model and performance enables data scientists and stakeholders to make informed decisions about model and deployment suitability, which drives potential improvements and optimizations to improve overall forecasting capabilities.

E. Tools and Techniques

The project uses a variety of tools to create a complete and efficient development environment. The project and backend are built using Flask, a lightweight and versatile web framework that facilitates seamless integration and communication. Jupyter Notebook serves as a dynamic and interactive computing environment that facilitates data exploration and analysis with a user-friendly interface. Visual Studio Code (Vs code) serves as the primary integrated development environment (IDE) that provides a robust platform for coding, debugging, and version control. The frontend of the project is designed using HTML, CSS and JavaScript, which ensures an interesting and responsive user interface.

IV. CONCLUSIONS

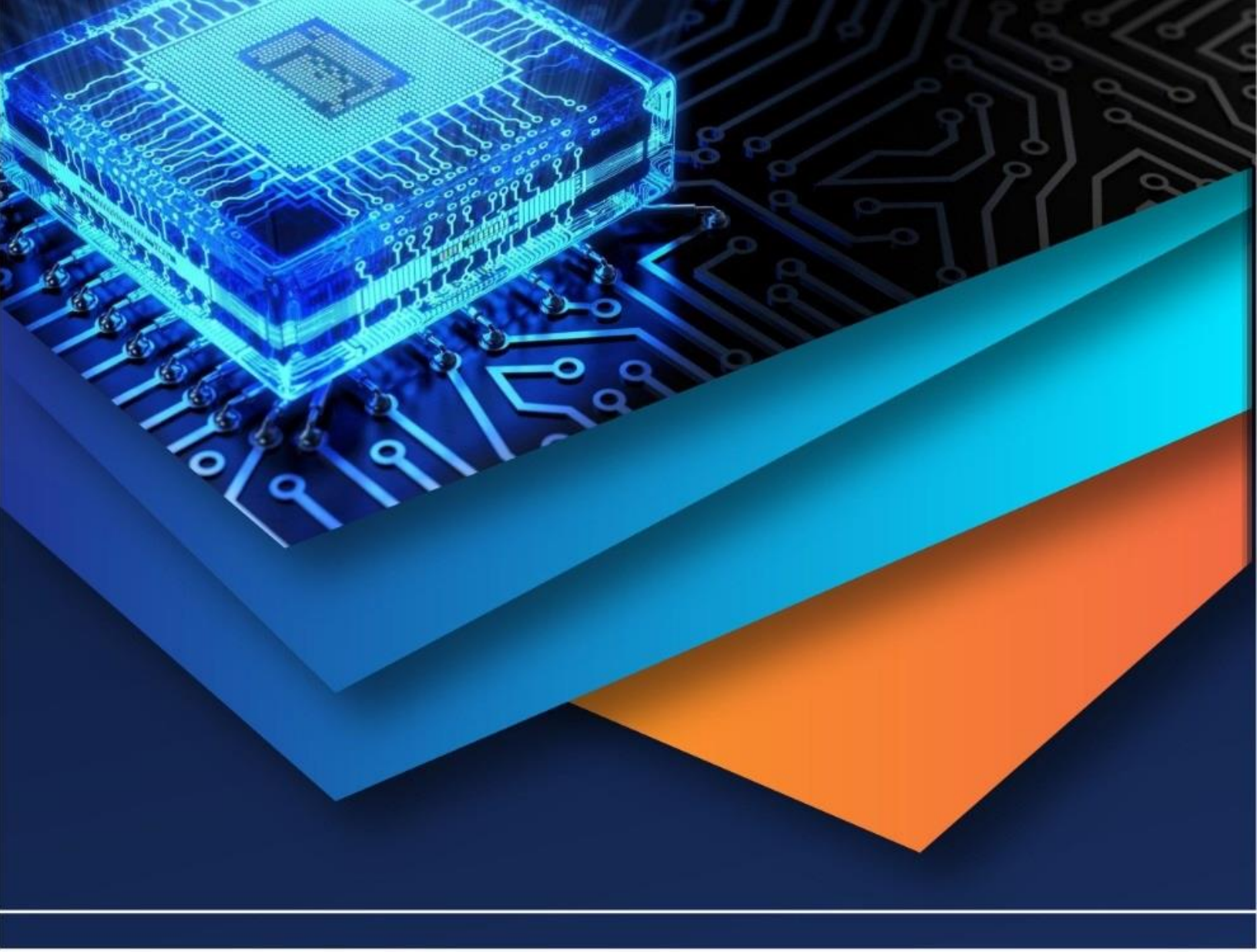
This project is a major step forward in the fight against the problems caused by PCOS, or polycystic ovarian syndrome. Our aim is to improve PCOS early identification and management by utilizing the capabilities of machine learning techniques and state-of-the-art healthcare technologies. By providing easily available resources for risk assessment and individualized counseling on PCOS management techniques, the project aims to empower individuals. We hope to offer insightful information through an easy-to-use interface, enabling consumers to make wise decisions regarding their health. Our ultimate goal is to support people who are coping with the challenges posed by PCOS by improving their overall quality of life and encouraging a proactive and supportive attitude toward health and wellbeing.

V. ACKNOWLEDGMENT

We would like to convey our sincere gratitude to Mrs. Swati Rajput for her great advice and mentoring, which were essential to the project's successful completion. Her knowledge, inspiration, and steadfast support helped to mold our concepts into a reliable and modern solution. We also sincerely thank our college for providing the tools and supportive atmosphere that enabled us to start this endeavor. The faculty's assistance and the stimulating classroom environment were very important to our academic progress. We would like to sincerely thank the developers of the tools and libraries that powered this project, as well as the open-source community. We have been continuously inspired by their spirit of collaboration and commitment to sharing information. The accomplishment of our project objectives was made possible in large part by the availability of these resources. In conclusion, we would like to express our sincere appreciation to our family members for their consistent help and understanding during the course of the project. Their support and tolerance were the cornerstones that kept us going when things got hard. We genuinely thank these people and organizations for their great assistance and encouragement, which helped us improve both academically and professionally. Their combined efforts were vital.

REFERENCES

[1] Dr R. Rekha, Ms. Srinithi V, "Machine learning for diagnosis of PCOS", IEEE Access, Sep 2023
[2] Sakthipriya Dhinakaran, Chandrakumar Thangavel, et al, "PCOS Perception analysis prediction using Machine learning algorithms", IEEE Access, 02 March 2023
[3] Akanksha Tanwar, Anima Jain, et al, "Accessible Polycystic Ovarian Syndrome Diagnosis Using Machine Learning", IEEE Access, May 2022
[4] Shazia Nasim, Mubarak Almutairi, et al, "A Novel Approach for Polycystic Ovary Syndrome Prediction using Machine Learning in Bioinformatics", IEEE Access, Sep 2022
[5] Kinjal Raut, Chaitrali Katkar, Prof. Dr. Mrs. Suhasini A. Itkar,"PCOS Detect using Machine Learning Algorithms", IRJET, Sep 2022
[6] Yasmine A. Abu Adla, Dalia G. Raydan, et al, "Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques",ICABME, 2021
[7] Dr Pooja Raundale, Harshil Kanakia, et al, "Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS", IEEE Access, June 2021
[8] Subrato Bharati, Prajoy Podder, et al, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms", IEEE Access, Sep 2022



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY