PCOS Detect using Machine Learning Algorithms

Kinjal Raut [1], Chaitrali Katkar [2], Prof. Dr. Mrs. Suhasini A. Itkar [3]

[1]Final Year Computer Engineering Student, PES Modern College of Engineering, Pune
[2]Final Year Computer Engineering Student, PES Modern College of Engineering, Pune
[3]Professor, Dept. of Computer Engineering, PES Modern College of Engineering, Pune, India

---

Abstract - *Polycystic ovary syndrome (PCOS), also known as polycystic ovarian syndrome, is hormonal endocrine disorder among women of reproductive age. Over five million women worldwide in their reproductive age are suffering from PCOS. The most common symptoms of this disorder may include missed periods, irregular periods, or very light periods, it affects in a way that ovaries become large or may contain many cysts, it can also cause excess body hair, including the chest, stomach, and hirsutism, can cause weight gain, especially around the abdomen, Acne or oily skin. The exact pathophysiology of PCOS is not yet known. This heterogenous disorder is characterized by the ovaries mainly. PCOS is a multifactorial and polygenic condition. Machine Learning is capable of "learning" features from very large amount through clinical practice to diagnose this disorder. This paper put forwards a solution to this problem which helps in early detection and prediction of PCOS treatment from an optimal and minimal set of parameters which have been statistically analyzed. The solution is built using machine learning algorithms such as Random Forest, Decision Tree, Support Vector Classifier, Logistic Regression, K-Nearest neighbors, XGBRF, CatBoost Classifier.*

Key Words: ***Machine Learning, Polycystic Ovary Syndrome, Random Forest, Decision Tree, Support Vector Classifier, K-Nearest Neighbours, Logistic Regression, K-Nearest Neighbours.***

1. INTRODUCTION

Technology is changing every outlook of our lives making remarkable transformations in the healthcare industry, nowadays technology and humans are working hand in hand. For example, robots performing surgeries once seemed a fiction but now they are performing critical and complex surgeries in hospitals.

Machine learning is a subclass of artificial intelligence, it helps the system learn, identify patterns of datasets, make logical decisions and performing digital analysis on digital information including words, numbers, images and clicks. Machine Learning applications mainly include image recognition, data prediction, Medical Diagnosis – Health Care and Clinical Care, etc. In this world of technology many advancements are taking place for detection of PCOS and Machine Learning algorithms are one of them.

PCOS is one of the most widely common endocrine disorders that affects 1 in 10 women of childbearing age. The exact prevalence of PCOS is not known but variable ranging from 2.2% to 26% globally. It was first detailed in 1935 by Stein and Leventhal as a syndrome manifested by hirsutism, and obesity associated with enlarged polycystic ovaries. Woman in reproductive age 15-40 experience hormonal imbalance, hence PCOS can happen at any age after puberty.

Hormones needed are progesterone, luteinizing hormone (LH), estrogen and follicle stimulating hormone (FSH). The common symptoms of PCOS are irregular menstrual cycle, too much hair, acne, weight gain, darkening of skin, skin tags. There is a high risk of first trimester miscarriage, in ovaries inappropriate growth of follicle can be prevented by detecting PCOS at an early stage. Hence detection of PCOS is important at primary stage. This paper focuses on prediction of PCOS.

The main work includes:

1. Selection of most important attributes using feature selection method from the dataset.
2. Applying/Performing machine learning algorithms on the selected features.
3. Comparing the performed algorithms in order to check accuracy.

1.1 Literature Review

Over 10 million young generation has been affected globally with 1 in every 4 four young women having PCOS. The disease is more common in urban population than rural because of the lifestyle. The increase in number of PCOS women is directly correlated with the sedentary lifestyle and lack of nutritional food, lack of exercise, weight gain and obesity.

Table -1: Summary of Literature review

AUTHORS	OBJECTIVES	RESEARCH DESIGN	RESULTS
Palak et al. [2012]	A method to automate PCOS based on clinical and metabolic markers.	Classification of features based on Bayesian and Logistic Regression.	Among the two compared models the best model built is Bayesian classifier with accuracy 93.93%.
Purnama et	Detection of	Three classi-	On C=40

al. [2015]	Follicles based on USG images by using the binary follicle images, feature extraction and segmentation	fication scenarios were designed Neural Network - LVQ, KNN - Euclidean distance, SVM - RBF kernel.	SVM-RBF kernel achieved 82% accuracy and on K=5 KNN achieved 78% accuracy.
Denny et al. [2019]	To overcome the time and cost involved in various clinical tests and ovary scanning.	PCOS features transformed with PCA used machine learning algorithms like KNN, SVM, RF, etc.	The best and accurate model for the PCOS detection came out Random Forest with 0.89 acc
Subrato et al. [2020]	Data driven diagnosis of PCOS using dataset on Kaggle repository.	Classifiers used are as follows gradient boosting, random forest, logistic regression, RFLR and methods applied are holdout and cross validation.	The best testing accuracy obtained is of RFLR 91.01%, recall value 90%.
Ning-Ning Xie et al. [2020]	To identify gene biomarkers and build diagnostic model	Computational method applied by combining two machine learning algorithms such as ANN, and Random Forest	A novel diagnostic model developed with accuracy of AUC: 0.7273 in microarray dataset and 0.6488 in RNA-seq dataset.
Priyanka et al. [2020]	Classification of PCOS will use physical symptoms and sonograms in which only the physical symptoms will be presented.	Used different algorithms like K-star, IB1 instance-based, locally weighted learning, Decision Table, M5 rules, Zero R, Random Forest and Random Tree to classify and find best model.	SVM-RBF kernel achieved 82% accuracy and on K=5 KNN achieved 78% accuracy.
Namarat Tanwani	A model is built using the causes and symptoms of PCOS as inputs and the output is predicted as presence or absence or PCOS.	Machine learning supervised classification algorithms used are K-NN and Logistic Regression.	The best accurate model built is Logistic Regression with accuracy 92%.
Madhumitha et al. [2021]	Ovary details are large range of follicles, type of cysts, follicle size, using image segmentation	Based on preprocessing and morphological operations SVM, KNN and Logistic Regression were used.	All three algorithms were combined and hybrid model were made and 0.98 accuracy were achieved
Pijush et al. [2021]	Detection and prevention of this disease as early as possible.	Used SMOTE and five other algorithms such as Logistic Regression, Random Forest, Decision Tree Support vector machine and K-NN together for early detection of PCOS.	The best model achieved accuracy, Training time: 97.11, F1 score: 0.010sec, Recall: 98%, Precision: 98% and AUROC: 95.6%
Khan Inan et al. [2021]	Conducting a probabilistic approach to select statistically relevant features which contribute to PCOS instances.	SMOTE, ENN and ANOVA Test, Chi-Square Test were used to identify important features. Classifiers such as XG Boost, SVM, KNN, NB, MLP, RF, AdaB were used.	G Boost outperformed all other classifiers with 0.96 accuracy and 0.98 Recall.

2. Methodology

Development of machine learning model to train the dataset is an important step for successful implementation. The dataset contains attributes such as I beta-HCG (mIU/mL), II beta-HCG (mIU/mL), AMH (ng/mL), Age(yrs), Weight (Kg),

Height (Cm), BMI, Blood Group, Pulse rate(bpm), RR (breaths/min), Fast food, Reg Exercise, BP_Systolic (mmHg), etc.

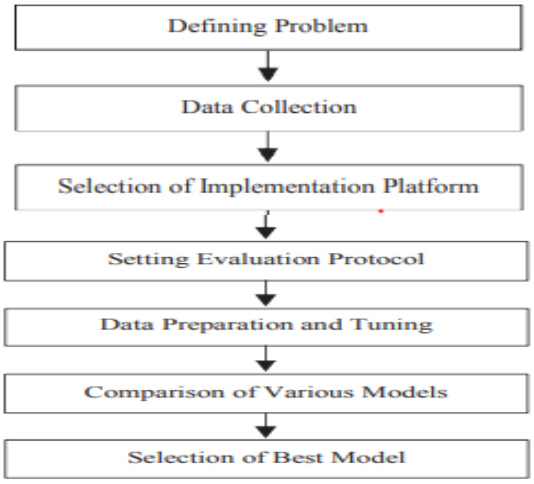

Fig -1: Block diagram of the system

1. Defining the Problem

   The first most important step in is to define the problem by including the inputs provided in the model and the expected output of the model.

2. Data Collection

   This is the crucial step of collecting data. How good the model will perform, the accuracy that we will get depends on the dataset. We can collect data from various platforms such as Kaggle, UCI Repository, BuzzFeed News, etc. We performed research on the dataset obtained on Kaggle named Polycystic ovary syndrome (PCOS).

3. Selection of implementation Platform

   For this machine learning implementation, the platform used is Jupyter Notebook, language used – Python.

4. Data Preparation

   When appropriate data is identified, the data should be shaped in order to train the model. The data obtained will be in csv format in python. Visualize the data and check the correlations between different characteristics. In this step checking for missing values or incomplete records, aggregation, augmentation, normalization, labelling, structured, unstructured and semi- structured data these activities are performed. This dataset contains the women patients in which they are suffering from PCOS. Further the steps to be performed in Data Preprocessing are:

i. Data Cleaning

It is the process of identifying the incorrect, incomplete or missing part of the data and then modifying, replacing or deleting them**.** In our paper the dataset was checked for missing values first by using the Pandas and Numpy2.

ii. Data Labeling

It is a method of identifying raw data i.e videos, images, text files, etc and add informative tags to provide context to increase the significance of machine learning model. The non-numerical are transformed into numerical values.

iii. Feature Selection

Feature Selection is an important step in which most relevant features are extracted from the dataset and then machine learning algorithms are applied for the better performance of model. It has a goal to find the best possible features for building the model ignoring the irrelevant details. Feature selection can be performed with common techniques including Filter methods, Wrapper methods and Embedded methods.

2.1 Modelling

When the data is completely cleaned and selected, it is ready to be processed by the algorithms. The algorithms used to create the model are Random Forest, Decision Tree, Support Vector Classifier, Logistic Regression, K-Nearest neighbors, XGBRF, CatBoost Classifier.

Random Forest

Random Forest is a kind of supervised machine learning algorithm used for both Classification and Regression. Its builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Decision Tree

Decision Tree is of type supervised learning algorithm, graphically represented for getting all possible solutions to a problem based on given conditions. It used CART algorithm which stands for Classification and Regression Tree algorithm.

Support Vector Classifier (SVC)

SVC is to fit the data provided returning a best fit that divides or categorizes the data. The data points are closer to the hyperplane and causes change in position and orientation of the given hyperplane.

Logistic Regression

Logistic Regression is a type of supervised Learning technique used for solving the classification problems. It is a

machine learning algorithm used for predicting the categorical dependent variable using a given set of independent variables and the cost function is limited between values 0 and 1.

K Nearest Neighbor (KNN)

K Nearest Neighbor also known as lazy learner algorithm is a supervised Learning algorithms used for both classification and regression. Instead of instantly learning the dataset, it first stores the dataset and then at the time of classification performs action on the given dataset.

XGBRF

XGBoost with Random Forest (XGBRF) is an ensemble method used for classification of PCOS. XGBoost is a gradient boosting algorithm and Random Forest is an example of bagging algorithm. XGBRF is a modified version of XGBoost classifier. The advantage of XGBRF is it is used to overcome the problem of over-fitting.

CatBoost Classifier

Categorical Boosting CatBoost or is an open-source boosting library used for regression and classification. It works with multiple categories of data, including audio, text and image including historical data. The technique used in this algorithm is to perform conversion from categorical values into numbers using different types of statistics on combinations of categorical features and combinations of categorical and numerical features.

Cross Validation

Cross-Validation in machine learning is a technique for validating the model efficiency in which model is trained using the subset of the dataset and after training the model is evaluated using the complementary subset of the dataset.

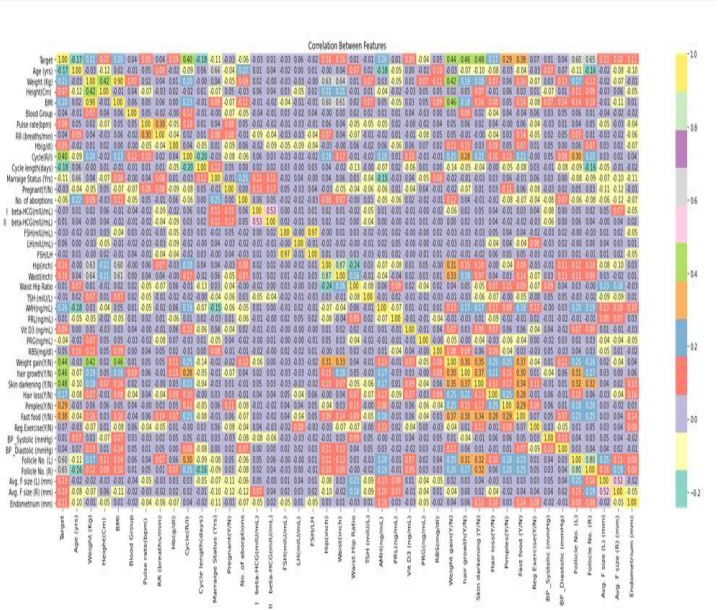

Fig -2: Correlation between Features

Results and Discussion

The experimentation is performed on the dataset using various machine learning algorithms. The main objective is to find most suitable algorithm for the classification of the dataset created. The algorithms used to construct the model are Decision Tree, SVC, Random Forest, Logistic Regression, K Nearest Neighbor, XGBRF and CatBoost Classifier.

Table -1: Accuracy of Different Classifier Models

Models	Accuracy
Decision Tree	82.79
SVC	69.05
Random Forest	89.42
Logistic Regression	83.32
K Nearest Neighbors	74.34
XGBRF	85.89
CatBoost Classifier	92.64

Fig -3: Accuracy of different Classifier Models

The accuracies obtained by different algorithms are: Decision Tree – 82.79%, SVC – 69.05%, Random Forest – 89.42%, Logistic Regression – 83.32%, K-Nearest Neighbors – 74.34%, XBRF – 85.89%, CatBoostClassifier – 92.64%.

Therefore, from the above results, conclusion is CatBoostClassifier has outperformed and obtained highest accuracy.

3. CONCLUSIONS

In this paper, Machine Learning model is successfully built and trained for early detection of PCOS. PCOS is one of the very common condition in women associated with psychological, reproductive and metabolic features. Some day-to-day activities to decrease the effects of PCOS are maintain a healthy weight, limit carbohydrates, be active, exercise daily and eat healthy food. The system in this paper helps in early detection of PCOS from an optimal and minimal set of parameters which have been statistically analyzed. Among the various algorithms used CatBoost Classifier is found superior in performance. This model can be used by doctors for early screening and diagnosing patients who are likely to develop this disorder. Therefore, with the use of various machine learning techniques we have built a model to detect PCOS at an early stage.

REFERENCES

[1] Palak Mehrotra, Jyotirmoy, Chatterjee, Chandan Chakraborty, "Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques", IEEE, 2012.

[2] Bedy Purnama, Untari Novia Wisesti, Adiwijaya, Fhira Nhita, Andini Gayatri, Titik Mutiah, "A Classification of Polycystic Ovary Syndrome Based on Follicle Detec-tion of Ultrasound Images, 2015 3rd International Conference on Information and Communication Tech- nology (ICoICT).

[3] 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).

[4] Subrato Bharati, Prajoy Podder, M. Rubaiyat Hossain Mondal, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms". 2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh.

[5] Ning-Ning Xie, Fang-Fang Wang, Jue Zhou, Chang Liu, Fan Qu, "Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network", Hindawi BioMed Research International Volume 2020.

[6] Priyanka R. Lele, Anuradha D. Thakare, "Comparative Analysis of Classifiers for Polycystic Ovary Syndrome Detection using Various Statistical Measures", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181: Vol. 9 Issue 03, March-2020.

[7] Namrata Tanwani, "Detecting PCOS using Machine Learning", IJMTES | International Journal of Modern Trends in Engineering and Science ISSN: 2348-3121, Volume:07 Issue:01 2020.

[8] J. Madhumitha, M. Kalaiyarasi, S. Sakthiya Ram, "Automated Polycystic Ovarian Syndrome Identification with Follicle Recognition", 2021 3rd International Conference on Signal Processing and Communication

[9] Pijush Dutta, Shobhandeb Paul, Madhurima Majumder, "An Efficient SMOTE Based Machine Learning classification for Prediction & Detection of PCOS", Research Square, November 8th, 2021.

[10] Muhammad Sakib Khan Inan, Rubaiath E Ulfath, Fahim Irfan Alam, Fateha Khanam Bappee, Rizwan Hasan, "Improved Sampling and Feature Selection to Support Extreme Gradient Boosting for PCOS Diagnosis.