VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi-560014, Karnataka

A PROJECT REPORT ON

"POLYCYSTIC OVARY SYNDROME DETECTION USING MACHINE LEARNING"

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE & ENGINEERING

Submitted By

ANUSHA B S (1SV20CS001)
ASHA (1SV20CS004)
KAVYA M (1SV20CS017)
MONIKA A (1SV20CS028)

Under the guidance of

Mrs. Sandhya K B.E., MTech

Assistant Professor, Dept. of CSE

SHRIDEVI EDUCATION

Department of Computer Science and Engineering

SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY
(Affiliated To Visvesvaraya Technological University)
Sira Road, Tumakuru – 572106, Karnataka.

2023-2024

Sri Shridevi Charitable Trust (R.)
SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY
Sira Road, Tumkur - 572 106, Karnataka, India.
Phone: 0816 - 2212629 | Principal: 0816 - 2212627, 9686114899 | Telefax: 0816 - 2212628
Email: info@shrideviengineering.org, principal@shrideviengineering.org | Website: www.shrideviengineering.org
(Approved by AICTE, New Delhi, Recognised by Govt. of Karnataka and Affiliated to Visvesvaraya Technological University, Belagavi)

ESTD:2002 — 20 YEARS — Academic Excellence

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that, the Project entitled "POLYCYSTIC OVARY SYNDROME DETECTION USING MACHINE LEARNING" has been successfullycarried out by ANUSHA B S [1SV20CS001], ASHA [1SV20CS004], KAVYA M[1SV20CS017], MONIKA A [1SV20CS028], in partial fulfillment for the award of **Bachelor of Engineering** in **Computer Science & Engineering** of the **Visvesvaraya Technological University, Belagavi** during the academic year **2023-24.** It is certified that all the corrections/suggestions indicated for internal assessments have been incorporated in the report. The Project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the Bachelor of Engineering Degree.

Sandhya K

Signature of Guide
Mrs. Sandhya K.,B.E., M.Tech.,
Assistant Professor,
Dept. of CSE,
SIET, Tumakuru.

29/5/24

Signature of H.O.D
Dr. Basavesha D.,B.E.,M.Tech.Phd
Associate Professor & HOD
Dept. of CSE,
SIET, Tumakuru.

Signature of Principal
Dr. Narendra Viswanath.,B.E., M.E., Ph.D., MIE, MISTE, MIWS, FIV.
Principal, SIET, Tumakuru

Name of the Examiners	Signature with date
1 Dr. Basavesha D	29/5/24
2 Rakesh S	29/May

Sri Shridevi Charitable Trust (R.)
SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY
Sira Road, Tumkur - 572 106, Karnataka, India.
Phone: 0816 - 2212629 | Principal: 0816 - 2212627, 9686114899 | Telefax: 0816 - 2212628
Email: info@shrideviengineering.org, principal@shrideviengineering.org | Website: www.shrideviengineering.org
(Approved by AICTE, New Delhi, Recognised by Govt. of Karnataka and Affiliated to Visvesvaraya Technological University, Belagavi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, **ANUSHA B S[1SV20CS001], ASHA[1SV20CS004], KAVYA M[1SV20CS017], MONIKA A[1SV20CS028]**, students of VIII semester **B.E** in Computer Science & Engineering, at Shridevi Institute of Engineering & Technology, Tumakuru, hereby declare that, the Project work-II entitled **"POLYCYSTIC OVARY SYNDROME DETECTION USING MACHINE LEARNING"**, embodies the report of our Project work carried out by our team under the guidance of **Mrs. Sandhya K, Assistant Professor, Department of CSE, SIET, Tumakuru** as partial fulfillment of requirements for the award of the degree in **Bachelor of Engineering in Computer Science & Engineering of Visvesvaraya Technological University, Belagavi**, during the academic year **2023-24.** The Project has been approved as it satisfies the academic requirements in respect to the Project work.

Place: Tumakuru

Date:

Student Names & Signatures

ANUSHA B S [1SV20CS001] ..... Anusha B.S.

ASHA [1SV20CS004] ..... Asha .....

KAVYA M [1SV20CS017] ..... Kavya M .....

MONIKA A [1SV20CS028] ..... Monika A.

ACKNOWLEDGEMENT

This Project will be incomplete without thanking the personalities responsible for this venture, which otherwise would not have become a reality.

We express our profound gratitude to **Dr. Narendra Viswanath,** Principal, S.I.E.T., for his moral support towards completing our Project work.

We would like to thank Head of Department **Dr. Basavesha D.,** Head, Department of CSE, SIET for providing all the support and facility.

We would like to thank my guide **Mrs. Sandhya K,** Assistant Professor, Department of CSE, SIET for her help, sharing her technical expertise and timely advice.

We would like to thank my Project Coordinators, **Dr. Girish L** Associate Professor, Head, Department of AI&DS and **Mrs. Rashmi N.,** Assistant Professor, Department of CSE, SIET for providing all the support and facilities.

We would like to express our sincere gratitude to all teaching and non-teaching staff of the Department of CSE for guiding us during the project work by giving valuable suggestion and encouragement.

By,

ANUSHA B S [1SV20CS001]

ASHA [1SV20CS001]

KAVYA M [1SV20CS017]

MONIKA A [1SV20CS028]

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is a critical hormonal disorder of women that significantly impacts life. In this new generation, women are more prone to PCOS. It is the cause of various problems, including infertility. Early detection of PCOS can reduce complexity. Therefore, an early and proper PCOS detection system is essential to minimize complications. Among all the detection techniques Machine Learning (ML) has an excellent performance in detection for its feature extraction capability. Therefore, considerable research has been carried out to detect PCOS using ML. Various ML approaches like Convolutional Neural Network, Support Vector Machine, K-Nearest-Neighbors, Logistic Regression, Decision Tree are used in detecting PCOS. A comprehensive analysis is carried out of how various ML approaches have been used in PCOS detection over the last few decades, and the techniques are discussed thoroughly. A complete examination was studied on differentdatasets used in PCOS detection. The performance of several algorithms is compared in quantitative and qualitative approaches. Finally, the significant difficulties andfuture research scopes are discussed to draw a conclusion.

CONTENTS

LIST OF FIGURES

Chapter: 1

INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is one of the most common hormonal disorders in women. It is an endocrine and metabolic disorder in premenopausal women. It is tough to recognize and treat PCOS. The correct cause of PCOS is not clear yet, but there is a close association with family history and hereditary qualities, hormones that are expanded during our advancement within the womb sometime recently birth, and way of life or environment.

It is commonly related to raised levels of two male hormones within the body. This hormone causes their body to skip menstrual periods and makes it harder for them to urge pregnant. Women with PCOS create a higher-than-normal number of male hormones. The side effects of PCOS are shown in numerous distinctive ways. A few women will have minor or gentle side effects, while others will have several serious side effects. Mood changes, sadness, uneasiness, and low self-esteem is some of the side effects of PCOS. Indications can, moreover, alter at distinctive stages of a woman's life.

Some of the primary symptoms of PCOS are no period or delayed period, immature ovarian eggs that do not ovulate, different 'cysts' on the ovaries, trouble in getting pregnant, excessive hair development often on the upper lip, chest, back, or buttocks, weight gain, thinning hair and hair misfortune from the head, oily skin or acne are some of the symptoms . Obesity, skin darkening, and skin pigmentation are more symptoms of PCOS. The effect of PCOS includes type-2 diabetes, cardiovascular disease, sleep apnea, mellitus, and trimester miscarriage.

- ✓ In addition to fertility impairment, a woman with PCOS may have some of the following symptoms and findings:
- Irregular or no menstrual periods in women of reproductive age (ovulatory dysfunction)
- Acne
- Weight gain
- Excess hair growth on the face and body (hirsutism)
- Thinning scalp hair

- Ovarian cysts (polycystic ovarian morphology)
- Mental health problems.

Women with PCOS are often resistant to the biological effects of insulin and, as a consequence, may have high insulin levels. Women with PCOS are at risk for type 2 diabetes, high cholesterol, and high blood pressure. Obesity also appears to worsen the condition. The degree of obesity may vary by ethnicity. however, this estimate does not include treatment of the serious conditions associated with PCOS.

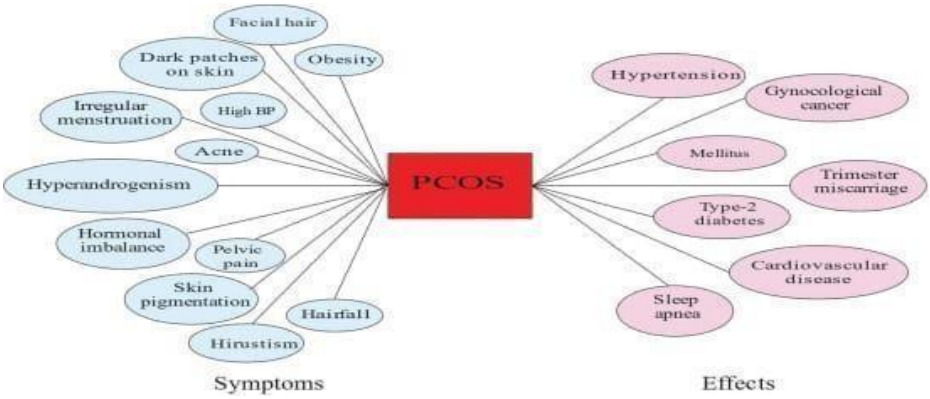


Figure 1: Symptoms and some basic effects that are found in women affected with PCOS. The left elements of PCOS are symptoms, and the right ones are effects of PCOS.

1.1 Overview

The primary need is to distinguish between PCOS and non-PCOS and treat- PCOS as early as possible. It requires a few tests and imaging methods as conceivable since the circumstance caused ovary disorder, which increases the chance of pregnancy complications, obstetric tumors, and mental trouble. Although much research was conducted to analyze PCOS utilizing different ML calculations, there is still a need for change in terms of exactness and precision based on medical information. Most women are unaware of their regenerative organs and the issues related to them. It causes infertility, uterus tumors, and closes with cancer. Parcels of therapeutic tests and time can discourage a PCOS-influenced woman from curing herself legitimately. Fig. represents the rate of PCOS affecting people around the world. It is seen that Asian women are infected most with PCOS, the rate is almost 31.3%, and White Americans are less infected with

4.8%. African Americans are infected at 6.8%, and Spanish women at 6.8%. This figure identifies race and genetic as two major elements behind PCOS in women.

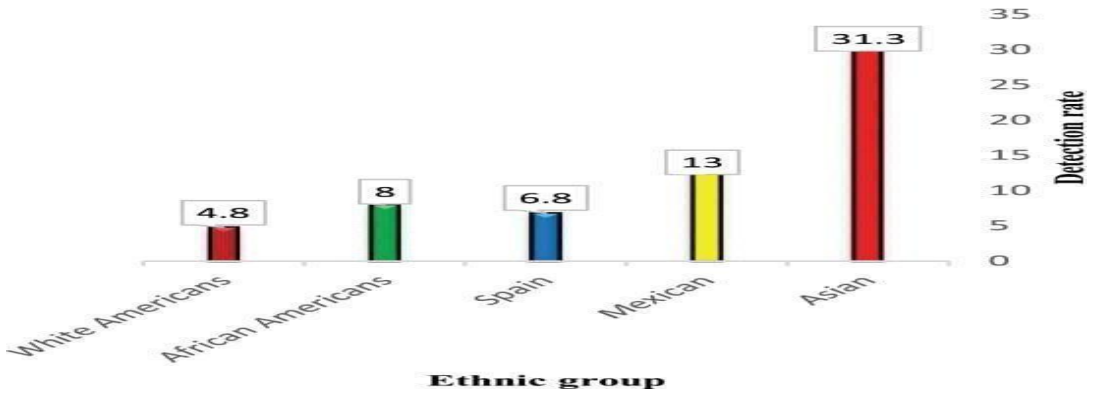


Figure 2: PCOS detection rate among women from different ethnicgroups around the world.

Innovation and human beings working together can create a way for better services in terms of health care. Machine learning a subsection of artificial intelligence where it gives the system a potential to learn and enhance automatically irrespective of being programmed clearly. It primarily focuses on developing new machine learning algorithms which give access to given datasets and use the data for study and research purposes of the openwork. Machine learning applications accompany huge transformation mainly in industries like health that involve diagnosis, image recognition, identification and prediction of data.

Polycystic ovary syndrome is an endocrine medical disorder which affects mainly women's throughout their period of adolescence. It was first narrated by Leventhal and stein in 1935. Women which are affected by polycystic ovary syndrome agonize from the imbalance of hormone level. It causes critical health problems such as irregular menstrual periods and problems related to getting pregnant. A woman while bearing a child faces hormonal irregularity in the middle of 15yr- 40yr age group. According to a research, the vulnerability of PCOS is 4.8% in white Americans, in African Americans it is 8% , in spain 6.8% and in Asia it is 31.3% . Women having PCOS suffer from diseases like hypertension, cardiovascular disease, type 2 diabetes, obesity, gynecological cancer, hazardous pregnancy and Mellitus.

Symptoms for PCOS are like acne problem, high blood pressure, irregular menstruation, increase in body weight, increase in androgen hormone levels etc. we examine PCOS as the main reason for infertility as it hold back the real evolution of follicle which anatomize the maturity of ovaries [Prapty and Shitu, 2020]. Recent research shows the high risk of first Trimester miscarriage.12-21% women in their reproductive age suffer from PCOS and out of them, 70% remains undetected. This disease can be treated by taking medication prescribed by doctors and changes in lifestyle habits. Medication includes birth control pills, diabetes tablets, medicine for anti androgen, fertility and ultrasound scan. Diagnosis of PCOS is done by barring of immaterial symptoms or test outcomes, mostly because of uneducated composite pathomechanism. These various symptoms force doctors to do a heavy number of clinical tests outcomes and irrelevant radio-logical imaging course of action.

The reproductive system of women wholly depend on unmatched hormones and required to be balanced for the processes which are needed for conception, ovulation and forming of a child in women's womb. Four hormones are needed namely progesterone, luteinizing hormone(LH), estrogen and follicle stimulating hormone(FSH). FSH and LH hormones are generated from Pituitary gland while as inside ovaries progesterone and estrogen are produced. For a good balancedreproductive system of women both progesterone and estrogen are very well important. Women with PCOS will have risks like sleep apnea, infertility, abnormal uterine bleeding, high cholesterol, elevated lipids, nonalcoholic fatty liver, liver disease, depression and anxiety, high blood pressure, metabolic syndrome, Miscarriages and Cardiac risks. PCOS can be predicted by symptoms like unwanted or excess growth of hairs in the body or face, Amenorrhea in 30 to 40% of women, increase in weight around the waist, before period swollen of breasts, during periods Neuralgic pain occurs, Hysteria, Itchy vagina and vulva and Cysts on ovaries.

According to the doctors, women having cysts in ovaries is not one of the main reason or parameter for diagnosis of PCOS. Studies suggest that 30-70% women having PCOS suffers mostly from obesity and it proves that there is a bidirectional relationship between PCOS and obesity. Although highly secretion of androgens, irregular menses and huge number of cysts in the ovary are declared as primary criteria for detection of PCOS. Studies show that these clinical features and can be used as important parameters for the early detection of PCOS.

The outcomes performed well than other methods used in and focusing on the success of this state of art approach, the primary goal is to implement these two algorithms on PCOS dataset for the detection of PCOS and evaluate the model on various metrics. It is still unknown that which algorithm will perform better and which algorithm will give best results in order to detect the PCOS.The vital challenge is to select the best attributes like which one is the crucial attribute for detection of PCOS in terms of FSH, LH, AMH, BMI, weight gain, cycle, Follicle NO, cycle length days, etc. Hence, we will implement the machine learning classifier algorithms, i.e. Hybrid Extreme Gradient Boosting with Random Forest (XGBRF) ensemble method and CatBoost Model, which includes physical and clinical parameters to detect the PCOS.

1.2 Problem Statement

- One of the reasons that PCOS can be hard to diagnose is that many of the symptoms it causes can also be caused by other conditions.
- Irregular or heavy menstrual bleeding, for example, can be caused by bleeding disorders, polyps, some medications, or uterine fibroids, along with several additional medical problems.
- There's no universal test for PCOS, which is another reason it can be difficult for women with this condition to get properly diagnosed.
- A physical exam, patient history, the symptoms you're having, an ultrasound, a pelvic exam, and blood test results can all be used together to make a better determination of whether you have PCOS.
- Still, not every doctor will make that diagnosis, especially if you don't have the most common symptoms of it, or if you have other medical issues that could be the cause of the problems you're experiencing.

1.3 Objectives

- To do gap analysis.
- To build a model which detects early and proper detection of PCOS.
- To compare performance of our model with existing system.
- To develop improved accuracy that lead to earlier, more accurate diagnosis and better overall outcomes for PCOS patients.
- To develop ML-based methods can identify patterns in medical data, such as hormone levels, to distinguish PCOS patients from those without the disorder. This improved accuracy could lead to earlier, more accurate diagnoses and better overall outcomes for PCOS patients
- To build the relationship between genetic and environmental factors that may explain PCOS by our model. We will use information collected during visits to our PCOS clinic and the family history of PCOS patients to learn about the causes and effects of PCOS.

1.4 Scope of Project

- Main aim is to use our model at hospitals to detect PCOS disorder.
- ML-based methods can identify patterns in medical data, such as hormone levels, to distinguish PCOS patients from those without the disorder.
- This provide a high diagnostic and classification performance in detecting PCOS, thereby providing an avenue for early diagnose of this disorder.
- To predict PCOS based on large at-risk population.
- This approach may guide early detection of PCOS within EHR-interfaced populations to facilitate counseling and interventions that may reduce long-term health consequences.
- To publish our project in Karnataka State Council for Science and Technology (KSCST).

1.5 Outcomes

- Early diagnosis enables timely intervention and management of symptoms.
- Improved understanding of hormonal imbalances and their effects on the body.
- Facilitates personalized treatment plans tailored to individual needs.
- Reduces the risk of long-term complications such as infertility, type 2 diabetes, and heart disease.
- Enhances quality of life through symptom management and lifestyle adjustments.
- Provides opportunities for fertility preservation and assisted reproductive techniques for those planning to conceive.
- Enables ongoing monitoring and adjustments to treatment plans as needed for optimal health outcomes.

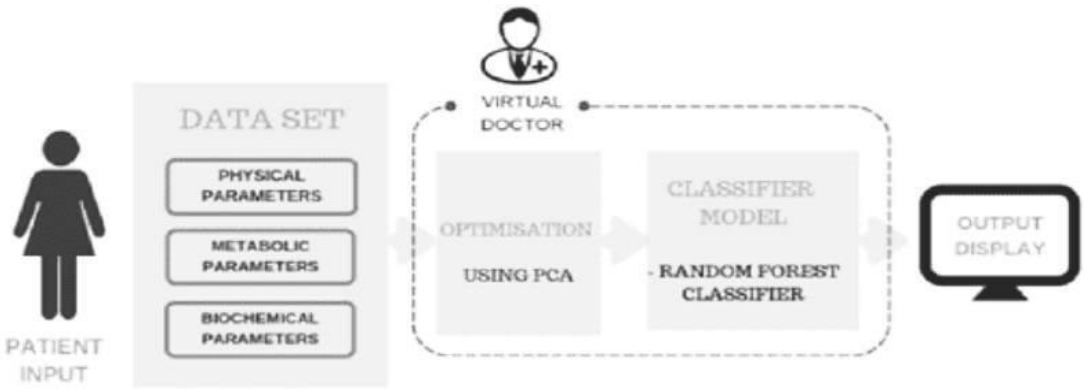


Figure 3: OUTCOME OF POLYCYSTIC OVARY SYNDROME DETECTION BY USING MACHINE LEARNING

Chapter 2:

LITERATURE SURVEY

Polycystic Ovary Syndrome (PCOS) is a medical condition which causes hormonal disorder in women in their childbearing years. Our proposed system helps in early detection and prediction of PCOS treatment from an optimal and minimal set of parameters. The Random Forest Classifier was found to be the most reliable and most accurate among 4 others with accuracy being 90.9%.

This section discusses some of the primary obstacles and difficulties associated with PCOS detection in previous studies to give a roadmap for researchers to examine where the emphasis should be put. Also, future directions are also discussed. With a figure the related difficulties for this work and their probable solutions are discussed. Fig. 16 demonstrates the challenges and their possible solutions. The clear representation of this figure will help the researchers in finding research gaps and conducting future works.

A. INFERIORITY OF STANDARD DATASET

Although a few effective data sets are available, there are some limitations. For example, the dataset available for PCOS detection is exceedingly small, and the datasets are not diverse. Most of the datasets are custom made. The custom datasets are very small. On the other hand, the number of datasets available on Kaggle are very few.ML works best when the trained and tested dataset is huge, as the model can learn and extract the features well. A large dataset that is broad in perspective and neutral to a particular geography is required. Moreover, a dataset should include women of various ages so that variety is included. If the dataset is not significant and standard, the tested result will not be accurate.

B. IMBALANCED DATASET

A balanced dataset has even types of observations for all classes. The existing datasets are effective, but they are not balanced. One class has a high number of observations, and

Our research study concludes that the dataset features prolactin (PRL), blood pressure systolic blood pressure diastolic, thyroid stimulating hormone (TSH), relative risk (RRbreaths), and pregnancy are the most prominent factors having high involvement in PCOS prediction. The study limitations and in future work, we will enhance the dataset by collecting more data on PCOS-related patients and applying data balancing techniques

This research provided a descriptive and conceptual assessment of all known PCOS detection techniques, focusing on ML. The method of existing algorithms was provided, as well their aspects, effectiveness, analysis methodology, and outputs. The flaws of existing algorithms were discussed, as well as potential problems. Though a significant amount ofresearch has been conducted to establish an efficient PCOS detection model, some problems remained unsolved. This paper detected the shortcomings, namely the small number of datasets, imbalance dataset, detection rate, not including more clustering approaches and so on.

In future work, we wish to work on a larger dataset having balanced data. Also, much CNN- based optimization needs to be conducted. We would like to use other clustering approaches like DBSCAN, OPTICS rather than K-means. This article will open a new window for the research community to understand the existing ML-based PCOS detection algorithms. Byunderstanding and analyzing the lackings and future scopes in this field, researchers will be able to develop new approaches to solve this problem.

In future work, we would like to study the new PCOS detection algorithms and apply these algorithms to a standard dataset to analyze the performance. We would like to use other clustering algorithms and understand the success rate.

This approach will make it easier tounderstand the major differences between the new and the old algorithms of ML for PCOS detection. the other class has a low number of observations. Due to this characteristic, the result is much affected. For this reason, no absolute effect is found. Various preprocessing techniques like PCA and SMOTE can be used to make the dataset balanced.

C. NOISE IN ULTRASOUND IMAGE

To run CNN model, the images must be clear enough, but ultrasound images are prone to speckle noise, salt and pepper noise, and many other noises. These noises must be removed to get the perfect result. As noise is not removed, then the accuracy rate for some CNN models are low.

As a result, identification of cysts is not perfect. But in the maximum existing literature, the noise from images is not removed. Therefore, for noise reduction dilation, grayscale, etc., these types of methods need to be used.

D. DETECTION RATE

If researchers want to make PCOS detection automatic and include ML-based techniques in the medical sector, then the detection rate must be 100% so that people can trust machine-based detection instead of manual detection. But in the previous works, detection rate is not perfect. Most of the detection rate is below 98%. This 2% accuracy must be increased by training the model more and more.

E. NOT INCLUDING OBJECT DETECTION

Object detection algorithms have brought a revolutionary change in the field of computer vision. This algorithm has the capability to object localization and object classification at the same time. As a result, object detection algorithms are faster than traditional approaches. Therefore, using object detection algorithms in detecting PCOS will be very beneficial. The existing literature basically focuses on classifier-based algorithms. But in thissector, YOLO, Fast RCNN, or this type of object detection algorithms can be used for detecting cysts from ultrasound images.

2.1 Machine Learning(ML):

ML is a growing field increasing area of processing techniques that targets to mimic human intelligence by learning from their surroundings. ML can be divided into two basic parts classification and clustering.

Machine learning can be broadly classified into four major types:

1. Supervised Machine Learning

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

2. Unsupervised Machine Learning

Unsupervised learning is a type of machine learning that learns from unlabeled data. This means that the data does not have any pre-existing labels or categories. The goal of unsupervised learning is to discover patterns and relationships in the data without any explicit guidance.

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without

any prior training of data. Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

You can use unsupervised learning to examine the animal data that has been gathered and distinguish between several groups according to the traits and actions of the animals. These groupings might correspond to various animal species, providing you to categorize the creatures without depending on labels that already exist.

3. Semi-supervised Machine Learning

Semi-supervised learning is a type of machine learning that falls in between supervised and unsupervised learning. It is a method that uses a small amount of labeled data and a large amount of unlabeled data to train a model. The goal of semi-supervised learning is to learn a function that can accurately predict the output variable based on the input variables, similar to supervised learning. However, unlike supervised learning, the algorithm is trained on a dataset that contains both labeled and unlabeled data.

Semi-supervised learning is particularly useful when there is a large amount of unlabeled data available, but it's too expensive or difficult to label all of it.

4. Reinforcement Learning

Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answerkey with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.

Reinforcement Learning (RL) is the science of decision making. It is about learning the optimal behavior in an environment to obtain maximum reward. In RL, the data is accumulated from machine learning systems that use a trial-and-error method. Data is not part of the input that we would find in supervised or unsupervised machine learning.

Machine Learning Techniques:

1. SVM [Support Vector Machine]:

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features. If thenumber of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number offeatures exceeds three.

Let's consider two independent variables x1, x2, and one dependent variable which is either a blue circle or a red circle.

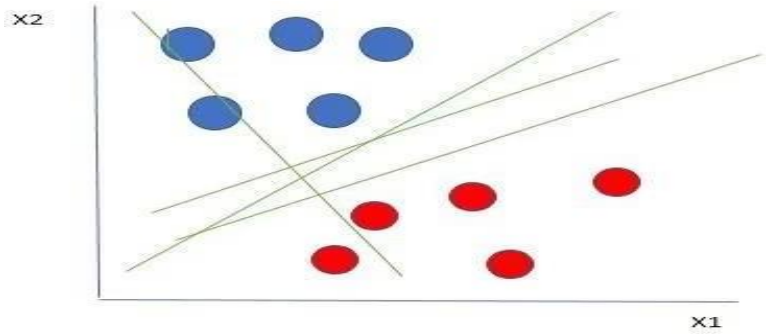


Figure 4: SVM Diagram

2. Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes:[2]

Decision nodes – typically represented by squares

Chance nodes – typically represented by circles

End nodes – typically represented by triangles

Decision trees are commonly used in operations research and operations management. If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm.[citation needed] Another use of decision trees is as a descriptive means for calculating conditional probabilities.

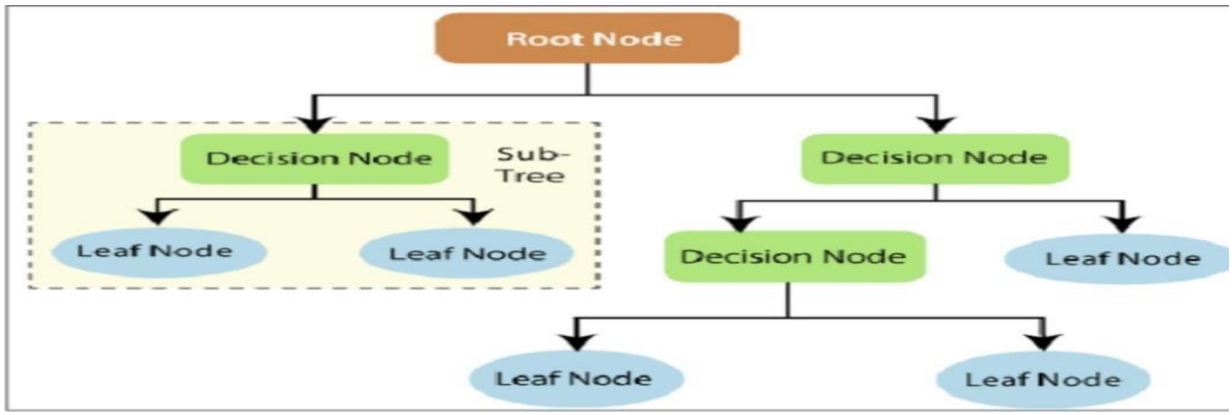


Figure 5: Decision Tree

3. KNN

The K-Nearest Neighbors (KNN) algorithm is a popular machine learning technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values.

During the training phase, the KNN algorithm stores the entire training dataset as a reference. When making predictions, it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.

Next, the algorithm identifies the K nearest neighbors to the input data point based on their distances. In the case of classification, the algorithm assigns the most common class label among the K neighbors as the predicted label for the input data point. For regression, it calculates the average or weighted average of the target values of the K neighbors to predict the value for the inputdata point.

The KNN algorithm is straightforward and easy to understand, making it a popular choice in various domains. However, its performance can be affected by the choice of K and the distance metric, so careful parameter tuning is necessary for optimal results.

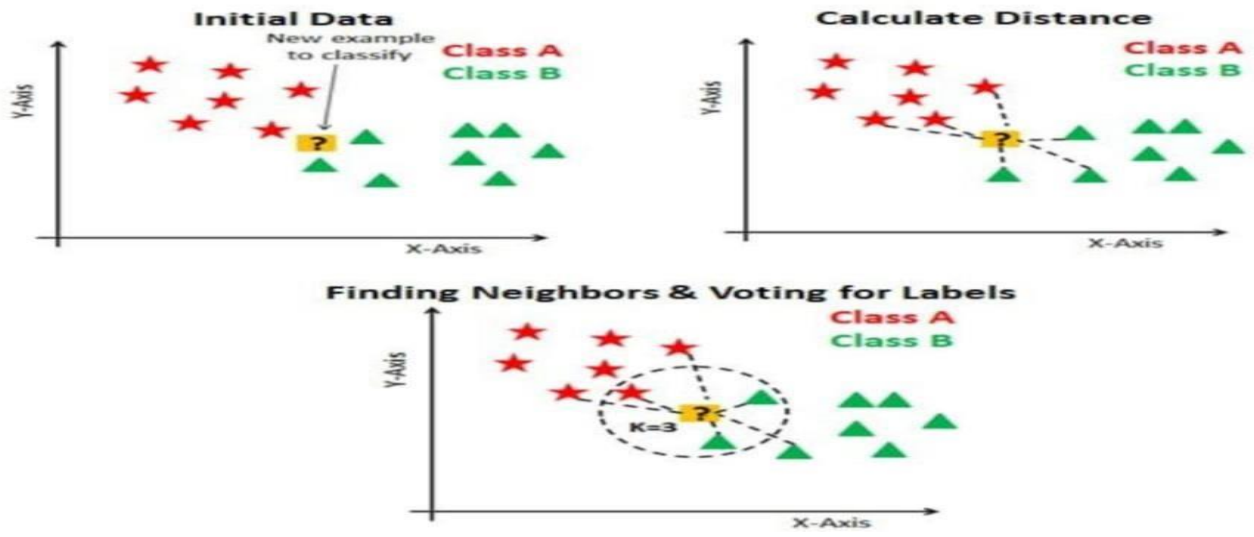


Figure 6: KNN Diagram

4. Logistic Regression

Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.

For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used for classification problems.

Key Points:

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.

It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

Logistic Function - Sigmoid Function

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.

The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

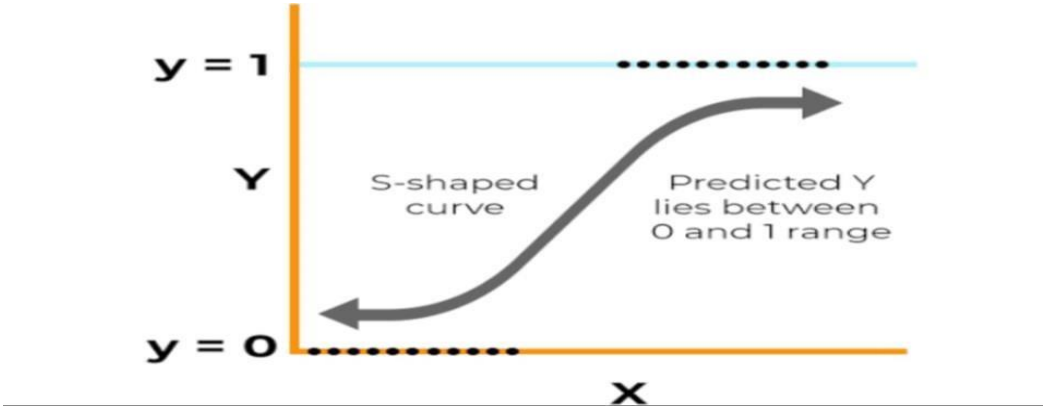


Figure 7: Logistic Regression

2.2 Related Work:

Detection of polycystic ovary syndrome has become a new topic for researchers since the last decade. Researchers have implemented various techniques to diagnose PCOS at an early stage. PCOS is an endocrine medical disorder with many diagnostic criteria because of its heterogenic manifestations. One of the primary diagnostic criteria includes examination of ovaries seen by ultrasound images in terms of number, size, and follicle distribution inside the ovary. This process includes manual tracing and follicle counting on the ultrasound images to decide PCOS. [Lawrence et al., 2007] presented a

novel method which automated the identification of PCOS. The algorithm included the follicle segmentation from ultrasound images, computing the features of the instinctively segmented follicles using follicle stereology, storing the follicle features as feature vectors, and at the last classification of feature vectors. They made two categories: one is PCO present and other is PCO absent. This automated tool saved a lot of time consumed with a manual tracing of follicles and calculating the width and length of each follicle.

Three classifiers were used, namely linear discriminant classifier(LDC), KNN and SVM. Results were very promising, as LDC achieved an accuracy of 0.92 , 0.91 by KNN and 0.91 by SVM, respectively. Overall, LDC outperformed SVM and KNN, but all three classifiers gave promising results. They reduced the risk of serious complications which can be caused by PCOS from delayed detection.

The development of various follicular cysts inside the ovary distinguishes this PCOS disorder.

First, the ultrasound image of the ovary is taken as input and filtered by using the adaptive morphological filter. Then adapted labeled watershed model is used to bring out the contours of targets. At last, clustering method is implemented to detect the expected follicle cysts. This investigation verifies the efficiency of the implemented automated scheme, which achieved the 0.84 accuracy. However, because of the high apposite of the PCOS, this automated scheme can not beimplemented to other various targets identification problems. There are two types of cervical mucus that are recognized, namely gestagenic and oestrogenic.

Goal was to detect the characteristics of crystallization and ultrastructure related to the cervical mucus in women which is suffering from PCOS and to contrast these characteristics with normal women. They took 10 samples of cervicalmucus from women, out of which 4 belongs to normal women and 6 were from women suffering from PCOS. Because of crystallization and ultra- structure, they characterized mucus. When the samples were taken, the levels got related to the type of mucus were progesterone and oestradiol. To consider mucus ultra-structure, they established differentiation between the women with PCOS and controlled women and anovulatory cycle of menses. These variations were obvious in the mesh and average mucus diameter of poses. In controlled women, Mucus crystallization showed the regular disposition of oestrogen like fern, hexagonal shape or rectilinear. While women having PCOS, unknown mucus crystallizations were established, apart from that patches of crystallization which resembles gestagenic-like mucus and oestrogen. then controlled women.

The PCOS can cause severe problems like anovulation and infertility. The criterion for PCOS detection includes metabolic and clinical parameters which are crucial for early pointer for this disease. [Mehrotra et al., 2011] described a new method which automated detection of PCOS because of these early markers. The model entails feature vector for-mulation based on the parameters of metabolic and clinical attributes.Further, significant statistical features for discerning between PCOS and normal groups are determined because of two sample t-test. Bayesian Classifier and Logistic Regression were implemented. This automated system acted as an assistant tool for the medical expertise which saved precious time in analysing the patients and thus reduced the delay therisk inof detection of PCOS.

Results were very promising, as Bayesian Classifier achieved accuracy of 0.93 and 0.91 were achieved by Logistic Regression. PCOS affects 10% women mainly when they are in reproductive

age and leads to infertility.proposed a new dosing system that is build on mathematical method for PCOS, a type of sterilities. This proposed system entitle us to initiate a new treatment for PCOS suited for sole PCOS patients. Performance is based on computer simulations. It produce sufficient LH surges unusual cycle by speed gradient model. If this system will be applied, it will be possible to give this dosage and helps in minimizing the side effects caused by PCOS especially pregnancy of women.

Described an automated scheme where the diagnosis of pathognomonic pattern and follicle arrangement is proposed to control this problem. The data is collected from GDIFR (GD Institute for Fertility Research) located in Kolkata. Patients from age group 25-35years diagnosed from PCOS are included. Ultrasonographer were used to get the ultrasounds of the patients by using the 7MHz transducer (General Electricals, Milwaukee, USA) and later verification were done by Gynaecologist. The imaging were pre-processed as per the methodology used. First, ultrasound image were pre-processed by the approach of multiscale morphological for contrast enhancement. Thenthresholding scaline is used for the extraction of follicle contours. The findings are then compared with the manual selection results to verify the potency of scheme.

PCOS disorder only affect women's health and it can be treated either by medication or surgery. Manual examination of PCOS detection frequently produce errors.described an algorithms which is capable of identifying cysts from the ultrasound images of ovaries and of alter between the kind of cysts. 25 digital recording with ultrasound images were taken for the model. Images were provided by MD. Barakat who is a gynaecologist. Images pre-processing were implemented in which images were converted into grey-scale and contrast enhancement were performed as well which are totally based on operation of morphology.

After this feature extraction were used to characteristics follicles based on standard parameters. SVM classifier were used for evaluation and validated by using ROC. 0.90 accuracy were achieved which is very promising. Nowadays, PCOS is a common disorder seen in women's which is caused by the development of various follicles in ovary. [Sitheswaran and Malarkhodi, 2014] described an effective model for the computer aided detection of PCOS by using the ultrasound images of ovary. The model is identifying follicles by using the object growing algorithm.

Chapter 3:

System Analysis

3.1 Existing System

Various types of PCOS detection techniques, their parameters, and their structures are described in this section. PCOS detection has two major groups of parent detection techniques. One is the traditional detection process, and the other is ML-based detection techniques. In the traditional detection techniques, the normal and PCOS hormonal range are very important. In this section, a detailed description of this hormonal range is given inside a table for easy understanding. The structures of ML algorithms are given for clear understanding of the researchers. A figure is included that clearly points out all the detection techniques of PCOS. Also, a detailed figure is given regarding the traditional detection process of PCOS, how doctors utilize traditional methods for PCOS detection.

Traditional Methods:

1) Hormonal Test And Symtom Aggregation

For PCOS detection, some hormone levels are considered that are:

i. Luteinizing Hormone (LH) and Follicle Stimulating Hormone (FSH): These two are important hormones for the female body. LH concentrations rise to around 25-40 mlU/ml before 24 hours of ovulation. Often women experiencing PCOS have LH levels of around 18mlU/ml and FSH levels of approximately six mlU/ml. This was once thought to be an essential factor in PCOS diagnosis.

ii. Testosterone: Overall, testosterone consists of the sum of all testosterones in the human body, including free testosterone. This ranges from 6.0 to 86 ng/dl. The quantity of testosterone in your body that is unbound and active is referred to as free testosterone, and the quantity is normally between 0.7 and 3.6 pg/ml. Both total and free testosterone levels are frequently elevated in women with PCOS.

iii. DHEA-S: These levels in most PCOS women are greater than 200 ug/dl.

iv. Prolactin: Women with PCOS have an increased rate of prolactin, often between 25 and 40 ng/ml.

v. Estrogen: Many women having PCOS have normal estrogen levels about 25-75 pg/ml).

vi. Thyroid Stimulating Hormone (TSH): The level of TSH in PCOS women is normally (0.4-3.8 uIU/ml). describes the range of normal and PCOS affected hormones. This table is useful for researchers to understand traditional PCOS detection methods. Along with the hormone levels, various symptoms are considered for PCOS detection. Some symptoms of PCOS detection are obesity, irregular menstrual, excessive hair fall, an increase in male hormones, etc. Physicians check the hormone level and consider the symptoms for detecting PCOS manually. The hormonal test is the most expensive and is a lengthy process to detect PCOS. A set of questionnaires is used to detect the symptoms of PCOS, traditionally performed. Questionnaires could be used to detect clinically obvious PCOS within relatives of PCOS patients. Though interviewing with a written questionnaire can find a high majority of afflicted mothers, approximately 50% of sisters with PCOS remain unreported .

2) Manual Ultrasound Image

Doctors manually detect PCOS based on the number of cysts from the ultrasound image. The presence of more than 12 follicles measuring 29 mm in each ovary may indicate PCOS. Doctors manually count the cysts and detect PCOS. It is time-consuming. Moreover, errors can occur, for example, a mistake during counting, not considering any cyst by mistake, considering any lump as a cyst, etc. Except for the abdomen ultrasound, another ultrasound is available. It is transvaginal ultrasound. In this process, a lubricated probe is inserted inside the vagina of the female, and the doctor sees the inside situation of the organs through a monitor. They analyze the situation of the uterus and cervix and manually count follicles on the ovaries. They also measure the volume of ovaries to detect PCOS. But this process is not error-free. Doctors can make mistakes while

counting follicles in the ovaries in measuring the volume of ovaries. Most importantly, this is not a painless method, and many people do not prefer this to social and cultural barriers.

3.2 Proposed System

The creation of a suitable machine learning model-based diagnostic tool for PCOS requires a comparison of the performance of several current algorithms in our data set. The preparation of the model, which gives the research its framework, is the most crucial phase. Withthe use of a workflow diagram and the procedures required in creating an acceptable model and fine-tuning it to provide the optimal outcome are described below.

I. **Data Collection**: This first step is to gather the relevant data from KAGGLE Analyze it and see if the information gathered is appropriate for our project.

II. **Pre-processing**: The following step in pre-processing is to remove null values. Later, all categorical textual data is converted into binary numerical data to aid in prediction. It's a statistical technique for evenly spreading out the number of instances in your collection. It functions by exploring innovative instances from minority situations that have already occurred.

III. **Feature Selection Making**: a choice feature that are Extremely Randomized Trees or Extended tree classifier ensemble machine learning technique is used to choose the characteristics that are crucial for PCOS illness prediction. It is analogous to classifiers from random forests. When compared to random forest, it differs in the way decision trees areconstructed in the forest. Based on the values of feature importance, features are classified as best or relevant.

Predictions in regression are created by averaging the decision tree output, but predictions in classification are made by majority voting.

IV. **Applying Algorithms**: A. Random Forest Algorithm Random Forest Numerous single decision trees that make up Random Forest work together to form an ensemble. The class with the most votes will be the class that the random forest's trees predict, and that class will be the model's output. Any single constituent model will perform worse than an enormous number of relatively uncorrelated models (trees) acting as a single committee.

B. CNN (Convolution Neural Network)Machine learning includes convolutional neural networks, sometimes known as convnets or CNNs. Convolutional neural networks, often known as convnets or CNNs, are a kind of machine learning. It is one of several artificial neural networkmodels that are utilized for a variety of activities and data sets. A CNN is a type of network design used in deep learning algorithms for applications such as image recognition and pixel dataprocessing. Primarily used for image processing.

V. **Connection Part**: Web Application Flask which is a python open-source software is used to build a web application, it is just a micro-framework it does not provide any database or form validation. HTML and CSS are used with flask which renders html pages and displays the content on the web page. CSS provides specifications for placement of html elements on the webpage. Users can give various inputs like sugar level, specific gravity, albumin etc.

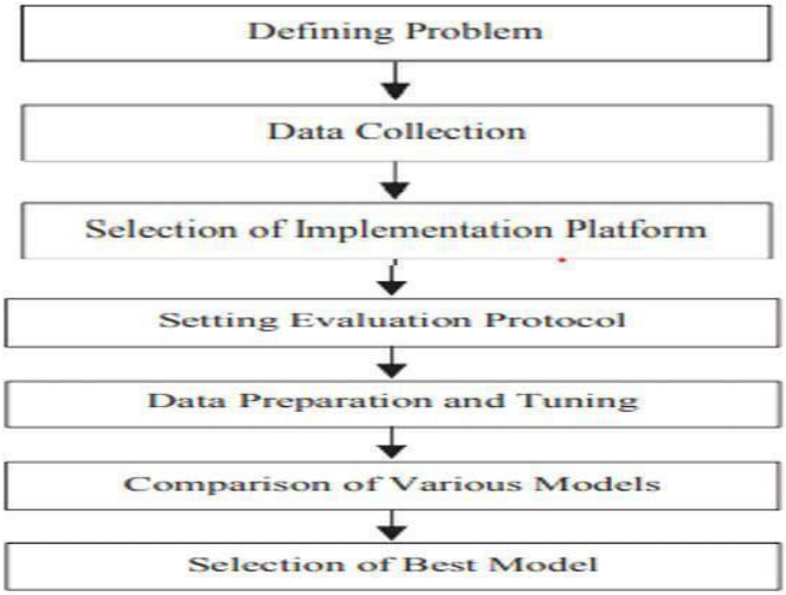


Figure 8: Block diagram of the system

Advantages of Proposed System:

1. Early Detection: Machine learning algorithms can analyze large amounts of data quickly and accurately, enabling early detection of PCOS symptoms before they become severe. This can lead to timely interventions and improved patient outcomes.

2. Improved Accuracy: Machine learning models can leverage complex patterns in data to make accurate predictions, potentially outperforming traditional diagnostic methods in terms of accuracy and efficiency.

3. Personalized Medicine: By analyzing individual patient data, machine learning models can tailor diagnoses and treatments to each patient's unique characteristics, optimizing outcomes and reducing the risk of misdiagnosis.

4. Efficiency: Automated PCOS detection using machine learning can streamline the diagnostic process, saving time for healthcare providers and patients and reducing the burden on healthcare systems.

5. Scalability: Machine learning models can be easily scaled to analyze large datasets from diverse populations, facilitating widespread adoption and improving the generalizability of diagnostic algorithms.

6. Continuous Improvement: Machine learning models can be trained and updated continuously with new data, allowing them to adapt to evolving patterns and trends in PCOS diagnosis and treatment.

7. Cost-Effectiveness: Automated PCOS detection using machine learning can potentially reduce healthcare costs by minimizing unnecessary tests & procedures through more targeted & efficient diagnostics.

Chapter 4:

IMPLEMENTATION

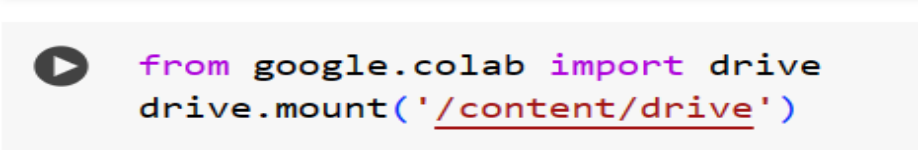
Our project aims to implement a machine learning model for the detection of Polycystic Ovary Syndrome (PCOS), a multifactorial endocrine disorder affecting women of reproductive age worldwide. While PCOS presents a wide range of symptoms, including irregular menstrual cycles, hormonal imbalances, and ovarian cysts, its diagnosis remains challenging due to its heterogeneous nature and overlapping features with other conditions. In response to this challenge, we have undertaken the implementation of a machine learning-based approach to facilitate accurate and timely PCOS detection.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
fromsklearn.metricsimport
accuracy_score,confusion_matrix,classification_report
from imblearn.over_sampling import SMOTE
```

Implementing a machine learning model for PCOS detection involves several key considerations, including dataset selection, feature engineering, model training, and evaluation. Our implementation builds upon existing research in PCOS diagnosis by leveraging a comprehensive dataset comprising clinical and demographic information collected from patients presenting with symptoms indicative of PCOS. This dataset serves as the foundation for training and testing our machine learning models, enabling us to develop a predictive model capable of discerning patterns and associations indicative of PCOS presence.

Central to our implementation is the selection and preprocessing of features relevant to PCOS diagnosis. By carefully curating the dataset and applying preprocessing techniques such as feature normalization, encoding categorical variables, and handling missing data, we ensure the quality and

integrity of the input data for our machine learning models. Feature selection methods are employed to identify the most informative predictors of PCOS, thereby enhancing the model's discriminative power and interpretability.



Mount at /content/drive

The heart of our implementation lies in the deployment and evaluation of machine learning algorithms for PCOS detection. We explore the performance of several popular algorithms, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), and k-Nearest Neighbors (KNN), assessing their ability to accurately classify patients as either PCOS-positive or PCOS-negative. Modeltraining involves optimizing algorithm-specific hyper parameters through techniques such as cross- validation, ensuring robustness and generalizability across diverse patient populations.

As part of our implementation, we conduct rigorous performance evaluation to assess the efficacy of each machine learning algorithm in PCOS detection. We employ standard evaluation metrics such as accuracy, precision, recall, and F1-score to quantitatively measure the models' performance andcompare their strengths and limitations. Our results provide valuable insights into the relative performance of different algorithms and inform the selection of the most suitable approach for PCOS detection in clinical practice.

In summary, our implementation endeavors to harness the power of machine learning to address the diagnostic challenges associated with PCOS. By developing and evaluating predictive models capable of discerning subtle patterns indicative of PCOS presence, we aim to empower healthcare practitioners with a valuable tool for early detection and intervention, ultimately improving patient outcomes and quality of care.

```
Df=pd.read_excel("../content/drive/MyDrive/PCOS_data_without_infertility.xlsx",sheet_name="Full_new")
df.head()
```

Dataset Description:

Our dataset comprises clinical and demographic information collected from patients undergoing evaluation for Polycystic Ovary Syndrome (PCOS) at a healthcare facility. Each observation in the dataset corresponds to a unique patient, with features encompassing a wide range of physiological, hormonal, and lifestyle variables potentially associated with PCOS development and progression.

The target variable, "PCOS (Y/N)", serves as the binary indicator of PCOS presence, with a value of 1 denoting positive PCOS status and 0 indicating absence. This variable forms the basis of our machine learning classification task, where our objective is to develop a predictive model capable of accurately discerning PCOS-positive individuals from those without the condition.

Key features included in the dataset are as follows:

1. Follicle No. (R/L):The number of ovarian follicles detected on the right and left ovaries, respectively, as determined by ultrasound examination. Follicle count abnormalities are common in individuals with PCOS and may serve as diagnostic markers.

2. Skin darkening (Y/N):Presence or absence of skin darkening, a dermatological manifestation associated with hormonal imbalances commonly observed in PCOS patients.

3. Hair growth (Y/N): Presence or absence of abnormal hair growth, particularly in androgen-sensitive areas such as the face, chest, and abdomen, indicative of hyperandrogenism, a hallmark feature of PCOS.

4. Weight gain (Y/N): Self-reported weight gain status, reflecting changes in body weight potentially associated with PCOS-related metabolic disturbances.

5. Cycle (R/I): Menstrual cycle regularity, categorized as regular (R) or irregular (I), with menstrual irregularities being a common symptom of PCOS.

6. Fast food (Y/N): Self-reported consumption of fast food, reflecting dietary habits that may influence metabolic health and PCOS risk.

7. Pimples (Y/N): Presence or absence of acne vulgaris, a dermatological manifestation associated with hormonal imbalances commonly observed in PCOS patients.

8. AMH (ng/mL): Anti-Müllerian hormone (AMH) levels, a biomarker of ovarian reserve and follicular development, often elevated in individuals with PCOS.

9. Weight (Kg), BMI, Cycle length (days): Anthropometric and menstrual cycle parameters potentially indicative of PCOS-related metabolic and reproductive abnormalities.

10. Hair loss (Y/N), Age (yrs), Waist (inch), Hip (inch): Additional clinical and demographic variables relevant to PCOS diagnosis and characterization.

The dataset also includes laboratory measurements such as hormone levels (e.g., LH, FSH, PRL), biochemical parameters (e.g., Hb, RBS), and imaging findings (e.g., ovarian morphology, endometrial thickness), which may provide further insights into the pathophysiology of PCOS and aid in diagnostic decision-making.

Overall, our dataset encompasses a comprehensive array of features representing various aspects of PCOS pathology and clinical presentation, laying the groundwork for the development of machine learning models for accurate and personalized PCOS detection and management.

```
correlation_with_target                                                 =
df.corr()['PCOS(Y/N)'].abs().sort_values(ascending=False)
print(correlation_with_target)
```

```
PCOS (Y/N)                1.000000
Follicle No. (R)          0.648327
Follicle No. (L)          0.603346
Skin darkening (Y/N)      0.475733
```

hair growth(Y/N)	0.464667
Weight gain(Y/N)	0.441047
Cycle(R/I)	0.401644
Fast food (Y/N)	0.377933
Pimples(Y/N)	0.286077
AMH(ng/mL)	0.263863
Weight (Kg)	0.211938
BMI	0.199534
Cycle length(days)	0.178480
Hair loss(Y/N)	0.172879
Age (yrs)	0.168513
Waist(inch)	0.164598
Hip(inch)	0.162297
Avg. F size (L) (mm)	0.132992
Marraige Status (Yrs)	0.112897
Endometrium (mm)	0.106648
Avg. F size (R) (mm)	0.097690
Pulse rate(bpm)	0.091821
Hb(g/dl)	0.087170
Vit D3 (ng/mL)	0.085494
Height(Cm)	0.068254
Reg.Exercise(Y/N)	0.065337
LH(mIU/mL)	0.063879
Sl. No	0.060998
Patient File No.	0.060998
No. of aborptions	0.057158
RBS(mg/dl)	0.048922
PRG(ng/mL)	0.043834
BP _Diastolic (mmHg)	0.038032
RR (breaths/min)	0.036928
Blood Group	0.036433
FSH(mIU/mL)	0.030319
I beta-HCG(mIU/mL)	0.027617
Pregnant(Y/N)	0.027565
FSH/LH	0.018336
II beta-HCG(mIU/mL)	0.013177
Waist:Hip Ratio	0.012386

```
TSH (mIU/L)                0.010140
BP _Systolic (mmHg)        0.007942
PRL(ng/mL)                 0.005143
Name: PCOS (Y/N), dtype: float64
```

Model Implementation:

In our implementation, we employed four distinct machine learning algorithms to develop predictive models for PCOS detection: Logistic Regression, Decision Tree, Support Vector Machine (SVM), and k-Nearest Neighbors (KNN). Each algorithm offers unique advantages and trade-offs in terms of interpretability, complexity, and robustness, making them well-suited for exploring the multifaceted nature of PCOS diagnosis.

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=
train_test_split(X,y,test_size=0.2,random_state=42)
```

1. Logistic Regression:

Algorithm Overview: Logistic Regression is a linear classification algorithm commonly used for binary classification tasks, such as PCOS detection. It estimates the probability that a given observation belongs to a particular class (PCOS-positive or PCOS-negative) based on a linear combination of input features.

Implementation: We trained a Logistic Regression model on our preprocessed dataset using gradient descent or other optimization techniques to learn the optimal coefficients for the linear decision boundary separating PCOS-positive and PCOS-negative instances.

Advantages: Logistic Regression offers simplicity, efficiency, and interpretability, making it a popular choice for healthcare applications where model transparency and clinical relevance are paramount.

```
logreg = LogisticRegression()
logreg.fit(x_train,y_train)
y_pred = logreg.predict(x_test)
accuracy = accuracy_score(y_test,y_pred)
conf_matrix = confusion_matrix(y_test,y_pred)
classification_rep = classification_report(y_test, y_pred)
```

2. Decision Tree:

Algorithm Overview: Decision Tree is a non-linear classification algorithm that partitions the feature space into hierarchical decision nodes based on the values of input features. It recursively splits the data into subsets, aiming to maximize information gain or minimize impurity at each node.

Implementation: We constructed a Decision Tree model to capture complex non-linear relationships between input features and PCOS status. The tree structure enables intuitive interpretation of feature importance and decision-making criteria.

Advantages: Decision Trees excel in capturing complex decision boundaries and interactions between features, making them well-suited for datasets with non-linear relationships and heterogeneous feature distributions, which are often observed in PCOS diagnosis.

```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(max_depth =4,random_state=42)
clf.fit(x_train,y_train
decision_pred = clf.predict(x_test)
decision_accuracy = accuracy_score(y_test,decision_pred)
decision_conf_matrix = confusion_matrix(y_test,decision_pred)
decision_classification_rep = classification_report(y_test, decision_pred)
```

3. Support Vector Machine (SVM):

Algorithm Overview: Support Vector Machine is a powerful classification algorithm that seeks to find the optimal hyperplane separating different classes while maximizing the margin between the nearest data points (support vectors). It can handle both linear and non-linear classification tasks through the use of appropriate kernel functions.

Implementation: We trained an SVM model using various kernel functions (e.g., linear, polynomial, radial basis function) to find the optimal decision boundary for PCOS classification. SVM's ability to handle high-dimensional feature spaces and non-linear relationships makes it well-suited for PCOS detection.

Advantages SVM offers flexibility in modeling complex decision boundaries and is less prone to overfitting, particularly in high-dimensional spaces. It can effectively capture intricate patterns in the data, enhancing its discriminative power in PCOS diagnosis.

```
from sklearn import svm
support=svm.SVC()
support.fit(x_train,y_train)
support_pred= support.predict(x_test)
support_accuracy = accuracy_score(y_test,support_pred)
support_conf_matrix = confusion_matrix(y_test,support_pred)
support_classification_rep = classification_report(y_test, support_pred)
```

4. k-Nearest Neighbors (KNN):

Algorithm Overview: k-Nearest Neighbors is a non-parametric classification algorithm that classifies new instances based on the majority vote of their k nearest neighbors in the feature space. It makes no explicit assumptions about the underlying data distribution and can accommodate arbitrary decision boundaries.

Implementation: We implemented a KNN classifier to predict PCOS status based on the similarity between input feature vectors and their nearest neighbors in the dataset. The choice of the number of neighbors (k) influences the model's bias-variance trade-off and generalization performance.

Advantages: KNN is simple, intuitive, and robust to noisy data, making it suitable for PCOS detection tasks where the underlying data distribution may be complex or poorly understood. It does not require explicit model training, allowing for straightforward implementation and interpretation.

In our implementation, each machine learning algorithm was trained and evaluated using standard practices such as cross-validation, hyperparameter tuning, and performance metrics assessment. By exploring multiple algorithms, we aimed to identify the most effective approach for PCOS detection based on our dataset characteristics and performance criteria. The selection of the final model was guided by considerations such as accuracy, interpretability, computational efficiency, and clinical relevance, ensuring a well-informed and evidence-based decision-making process in PCOS diagnosis and management.

```
from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=3)
neigh.fit(x_train,y_train)
neigh_pred=neigh.predict(x_test)
neigh_accuracy = accuracy_score(y_test,neigh_pred)
neigh_conf_matrix = confusion_matrix(y_test,neigh_pred)
neigh_classification_rep = classification_report(y_test, neigh_pred)
```

Module Evaluation:

In our endeavor to develop a robust and accurate model for Polycystic Ovary Syndrome (PCOS) detection using machine learning, we meticulously evaluated the performance of four distinct algorithms: Logistic Regression, Decision Tree, k-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The evaluation process aimed to assess each algorithm's effectiveness in correctly

classifying individuals as PCOS-positive or PCOS-negative based on their clinical and demographic features.

```
Accuracy: 0.8888888888888888
Confusion Matrix:
[[74  5]
 [ 7 22]]
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.94      0.92        79
           1       0.81      0.76      0.79        29

    accuracy                           0.89       108
   macro avg       0.86      0.85      0.86       108
weighted avg       0.89      0.89      0.89       108
```

Logistic Regression demonstrated the highest accuracy among the evaluated algorithms, correctly classifying 88.89% of instances. It exhibited balanced precision and recall rates for both PCOS- positive and PCOS-negative cases, indicating its effectiveness in making accurate predictions across different classes.

```
Accuracy: 0.8703703703703703
Confusion Matrix:
[[74  5]
 [ 9 20]]
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.94      0.91        79
           1       0.80      0.69      0.74        29

    accuracy                           0.87       108
   macro avg       0.85      0.81      0.83       108
weighted avg       0.87      0.87      0.87       108
```

Decision Tree achieved an accuracy of 87.04%, slightly lower than Logistic Regression. While it exhibited high precision and recall for PCOS-negative cases, it showed lower performance in correctly identifying PCOS-positive instances, indicating room for improvement in capturing positive cases.

```
Accuracy: 0.8333333333333334
Confusion Matrix:
[[72  7]
 [11 18]]
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.91      0.89        79
           1       0.72      0.62      0.67        29

    accuracy                           0.83       108
   macro avg       0.79      0.77      0.78       108
weighted avg       0.83      0.83      0.83       108
```

KNN achieved competitive performance with an accuracy of 83.33%. While it demonstrated high precision and recall for PCOS-negative cases, it exhibited challenges in correctly identifying PCOS-positive instances, suggesting the need for further refinement in capturing positive cases.

```
Accuracy: 0.7870370370370371
Confusion Matrix:
[[70  9]
 [14 15]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.89      0.86        79
           1       0.62      0.52      0.57        29

    accuracy                           0.79       108
   macro avg       0.73      0.70      0.71       108
weighted avg       0.78      0.79      0.78       108
```

SVM exhibited the lowest accuracy among the evaluated algorithms, achieving a score of 78.70%. While it demonstrated high precision and recall for PCOS-negative cases, it showed limitations in correctly identifying PCOS-positive instances, indicating the need for improvement in capturingpositive cases.

The module evaluation provided valuable insights into the performance of each algorithm for PCOS detection. Logistic Regression emerged as the top-performing algorithm, demonstrating the highest accuracy and balanced precision-recall trade-off. Decision Tree and KNN also showed competitive performance, while SVM exhibited relatively lower accuracy and predictive power. These insights guide the selection and optimization of machine learning models for PCOS detection, ensuring robust and reliable performance in real-world applications. Further research and experimentation may be warranted to explore ensemble methods, feature engineering techniques, and alternative algorithmic approaches to enhance the accuracy and robustness of PCOS detection models.

Chapter 5:

RESULTS AND ANALYSIS

Accuracy Analysis

The logistic regression model achieved the highest accuracy among the evaluated algorithms, with a score of 88.89%. This indicates that the model accurately predicted the PCOS status for approximately 89% of the instances in the dataset. A high accuracy score suggests that the model's predictions align closely with the actual PCOS status, demonstrating its effectiveness in distinguishingbetween PCOS-positive and PCOS-negative cases.

```
Accuracy: 0.8888888888888888
Confusion Matrix:
[[74  5]
 [ 7 22]]
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.94      0.92        79
           1       0.81      0.76      0.79        29

    accuracy                           0.89       108
   macro avg       0.86      0.85      0.86       108
weighted avg       0.89      0.89      0.89       108
```

The decision tree algorithm yielded an accuracy of 87.04%. While slightly lower than logistic regression, it still performed well in classifying PCOS cases. Decision trees are known for their ability to capture complex decision boundaries and interactions between features. Despite this, the model maintained a high level of accuracy, showcasing its suitability for PCOS detection tasks.

KNN achieved an accuracy of 83.33%, demonstrating competitive performance in PCOS detection. KNN relies on the similarity between instances in the feature space and assigns labels based on the majority class among nearest neighbors. Despite its simplicity, the model was able to accurately classify a significant portion of instances, underscoring its effectiveness as a classification algorithm for PCOS detection.SVM exhibited the lowest accuracy among the evaluated algorithms, with a score of 78.70%.

While SVMs are powerful classifiers capable of capturing complex relationships in high-dimensional spaces, the model's performance in PCOS detection fell short compared to other algorithms in this study. This suggests that the chosen SVM configuration or kernel may not be optimal for the given dataset, highlighting the importance of hyper parameter tuning and model selection in achieving better performance

Precision and Recall Analysis:

The logistic regression model achieved high precision and recall rates for both PCOS-positive and PCOS-negative cases. This indicates that the model not only minimized false positives but also effectively captured most of the true positive cases. The balanced precision and recall demonstrate the model's ability to make accurate predictions across different classes.

While decision tree achieved high precision for PCOS-negative cases, it exhibited lower precision and recall rates for PCOS-positive cases. This suggests that the model may struggle to correctly identify positive instances, potentially leading to missed diagnoses. Further analysis of the decision tree's structure and feature importance could provide insights into areas for improvement.

KNN demonstrated competitive precision and recall rates, albeit lower than logistic regression. The model exhibited high precision for PCOS-negative cases but showed limitations in correctly identifying PCOS-positive cases. This suggests that the model may need adjustments to better capture the underlying patterns in positive instances.

Chapter 6:

CONCLUSION

In this study, we investigated the efficacy of machine learning algorithms for Polycystic Ovary Syndrome (PCOS) detection using clinical and demographic features. Through comprehensive model evaluation, including logistic regression, decision tree, k-Nearest Neighbors (KNN), and Support Vector Machine (SVM), we gained valuable insights into their performance characteristics and suitability for PCOS diagnosis.

Our findings reveal that logistic regression emerged as the top-performing algorithm, achieving the highest accuracy and balanced precision-recall trade-off. The model demonstrated the ability to accurately classify PCOS-positive and PCOS-negative cases, highlighting its potential as an effective tool for PCOS detection.

While decision tree and k-Nearest Neighbors (KNN) also showed promise, further optimization may be necessary to improve their performance, particularly in correctly identifying PCOS-positive cases. Additionally, Support Vector Machine (SVM) exhibited suboptimal accuracy and may require alternative approaches or parameter tuning to enhance its effectiveness in PCOS detection.

Overall, our study underscores the importance of rigorous model evaluation and optimization in developing accurate and reliable PCOS detection models. Future research efforts should focus on exploring ensemble methods, feature engineering techniques, and larger datasets to further improve model performance and advance the field of PCOS diagnosis using machine learning.

In conclusion, our study contributes to the growing body of knowledge in PCOS diagnosis and underscores the potential of machine learning algorithms as valuable tools in clinical decision-making and patient care.

Future Scope:

Based on the results, logistic regression emerges as the top-performing algorithm for PCOS detection, achieving the highest accuracy and balanced precision-recall trade-off. Decision tree and k-Nearest Neighbors (KNN) also show promise but may require further optimization to enhance their performance, particularly in correctly identifying PCOS-positive cases. Support Vector Machine (SVM) exhibits the lowest accuracy and may benefit from additional tuning or alternative approaches to improve its effectiveness in PCOS detection.

Further research could focus on ensemble methods, feature engineering techniques, and alternative algorithmic approaches to enhance the accuracy and robustness of PCOS detection models. Additionally, larger and more diverse datasets could be explored to improve model generalization and real-world applicability. Continued efforts in model optimization and validation are essential to advance the field of PCOS diagnosis using machine learning

Chapter 7:

REFERENCE

- https://ieeexplore.ieee.org/document/10076750/
- https://ieeexplore.ieee.org/document/10403567/
- https://ieeexplore.ieee.org/document/10276245/
- https://www.geeksforgeeks.org/polycystic-ovary-syndrome-pcos/
- https://www.javatpoint.com/types-of- machine-learning
- https://www.researchgate.net/publication/348627784_PCOcare_PCOS_Detection_and_Prediction_using_Machine_Learning_Algorithms
- https://www.researchgate.net/publication/348627784_PCOcare_PCOS_Detection_and_Prediction_using_Machine_Learning_Algorithms
- https://www.researchgate.net/publication/348627784_PCOcare_PCOS_Detection_and_Prediction_using_Machine_Learning_Algorithms
- https://www.researchgate.net/publication/348627784_PCOcare_PCOS_Detection_and_Prediction_using_Machine_Learning_Algorithms