A PROJECT REPORT ON

DESIGN AND ANALYSIS OF A SYSTEM FOR PCOS DETECTION

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

BACHELOR OF ENGINEERING (COMPUTER ENGINEERING)

SUBMITTED BY

SAYALI DEODIKAR
SHRADDHA JADHAV
AISHWARYA JOSHI
SEJAL MUTAKEKAR

Under The Guidance of

Prof. Asma Shaikh

DEPARTMENT OF COMPUTER ENGINEERING
Marathwada Mitra Mandal's College of Engineering
Karvenagar, Pune-411052
Savitribai Phule Pune University

2022-23

CERTIFICATE

This is to certify that the project report entitled

"DESIGN AND ANALYSIS OF A SYSTEM FOR PCOS DETECTION"

Submitted by

SAYALI DEODIKAR
SHRADDHA JADHAV
AISHWARYA JOSHI
SEJAL MUTAKEKAR

are bonafide students of this institute and the work has been carried out by them under the supervision of **Prof. Asma Shaikh** and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University, for the award of the degree of **Bachelor of Engineering**(Computer Engineering).

Prof. Asma Shaikh
Internal Guide,
Dept. of Computer Engg.

Dr. K. S. Thakre
Head,
Dept. of Computer Engg.

Dr.V. N. Gohokar
I/C Principal,
Marathwada Mitra Mandal's College of Engineering, Karvenagar, Pune – 411052

Date:

Acknowledgment

We take this to express our deep sense of gratitude towards our esteemed guide Prof. Asma Shaikh for giving us this splendid opportunity to select and present this project and also providing facilities for successful completion.

I thank Dr.Kalpana Thakre, Head, Department of Computer Engineering, for opening the doors of the department towards the realization of the project, all the staff members, for their indispensable support, priceless suggestion and for most valuable time lent as and when required. With respect and gratitude, we would like to thank all the people, who have helped us directly or indirectly.

SAYALI DEODIKAR
SHRADDHA JADHAV
AISHWARYA JOSHI
SEJAL MUTAKEKAR

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is a disease that poses a serious threat to women's health in the 21st century. Although it occurs rarely, it can lead to permanent infertility and gynecological cancer. PCOS is characterized by the overproduction of a hormone called androgen. The traditional method for detecting PCOS involved manual examination of ultrasound images and follicle properties, which was time-consuming and prone to errors. To address this issue, a new system has been proposed to detect PCOS using advanced technology. The proposed system comprises two modules: the first module applies CNNs to ultrasound images, while the second module develops a semi-supervised learning-based PCOS test. By automating the PCOS detection process, the proposed system offers a faster, more accurate, and simpler alternative to the traditional method. It is expected to enhance the diagnosis and treatment of PCOS, thus improving women's health outcomes.

Technical Keywords

(1) Data Science

(2) Machine Learning

(3) Deep Learning

(4) Convolution Neural Networks

(5) Python Libraries

(6) Algorithms

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Overview

Polycystic ovary syndrome is a disorder involving a prolonged menstrual cycle. PCOS is a condition in the female body which causes multiple sacs in the ovaries. One in five that is almost 20 percent of the women population suffer from this syndrome. In most women suffering from PCOS missed or irregular menstrual periods, excess hair growth, acne, infertility, and weight gain are the common symptoms. Women suffering from PCOS may be at higher risk for diseases like type 2 diabetes, depression, high blood pressure, anxiety, heart problems, and endometrial cancer. The Diagnosis of PCOS is mostly done with three types- 1) Ultrasound test- This test is performed with sound waves over the ovaries to find out the size of the ovaries and if there are cysts in the ovaries. It also checks the thickness of the lining of the uterus.2) Self-diagnosis: Most women start to notice the major symptoms of PCOS like irregular periods, excess hair growth, and weight gain in the early stages, and with the help of a gynecologist it is confirmed. 3) Blood test- In the blood tests for PCOS they look for high levels of androgens and other hormones. It is checked if the reason behind symptoms is not other diseases like thyroid.

In most cases, doctors manually examine ultrasound images and conclude the affected ovary but are unable to find whether it is a simple cyst or PCOS. As cysts are very fine, doctors take time to diagnose PCOS with high accuracy. Along with that, many women suffer from conditions such as heavy bleeding, weight gain, excessive hair growth, etc. But due to the lack of knowledge, they don't understand that these are the symptoms of PCOS.

Algorithmic trading has transformed the normal relationship between investors and their market access agents in the trading of agents. Computer algorithms generate trading orders for individual tools without human intervention used internally by trading third-party firms. However, with

the help of new market access models, the purchasing side gained more control over the actual trading decision and order distribution processes and was empowered to develop and implement their own trading algorithms or use standard software solutions from independent software vendors. However, the sales side still offers most of the algorithmic trading tools to their customers. Using computer-generated automated computer algorithms has reduced all of the trading costs for investors, as there are no longer expensive private traders involved. As a result, Algorithmic trading has acquired significant stock markets in international financial markets in recent years.

1.2 Motivation

The exact prevalence of PCOS is not clear but 1 out of 10 women is affected, which is ranging between 2.5 percent to 25 percent. Not everyone comes up with this issue and prefers to keep this as personal suffering. They prefer keeping this as confidential information. In recent years, research has proved this to be harmful and suggests its early diagnosis. Hence, this makes it the need of an hour to make use of technology to detect PCOS and have a proper treatment to mitigate its effects and future consequences. To avoid this disorder being a base for many other complementary diseases, there is a need to diagnose this disease and gain a proper treatment. This may turn out to be the first successful step for completely healing this disorder.

1.3 Problem Definition

PCOS diagnosis for doctors might take a long time and needs high accuracy for better treatment decisions. The problem we are trying to solve here is to detect the PCOS based on ultrasound images using machine learning and deep learning based methods. A user-friendly environment can help women as well as doctors, easily detect the results within a few seconds. This system will enable them to upload images or enter the information to get the report. The Images will be classified by Convolution Neural Networks, which are considered to be the best option in recent times. Moreover, taking user inputs (Eg: a questionnaire related to symptoms) can help a system come to a conclusion. Not only will the women be able to detect if they are suffering from the syndrome, but also they will be introduced to the nearest possible hospitals, where they can have further treatment.

1.4 Project Scope and Limitations

Detection of PCOS involves a number of factors and tests such as blood test, genetic test, obesity etc. One of such tests is ultrasound test also known as sonography. This project is based on the images obtained by ultrasound test. In it we aim to consider more tests. Along with that the PCOS has several types. The scope of this project is limited to detection of PCOS in females. In future we will try to build a system for detailed analysis of PCOS like which type it is, what stage, what are its side effects etc. As sometimes females hesitate to go to the doctor for their problems so it is helpful for females to know whether they have PCOS or not by taking small quizzes or if they have sonography images by entering that they are able to detect the disease.

Chapter 2

LITERATURE SURVEY

2.1 Literature Review

Islam (2022) et. al [1], suggested an approach for the detection of PCOS using ovary ultra-sonography (USG) scans.
An extended machine-learning classification technique for PCOS prediction has been proposed on over 594 images for training and testing purposes. In it, the CNN algorithm with feature extraction technique is used for extracting features from an image then stacking ensemble machine learning technique using conventional models as base learners and bagging and boosting ensemble model as a meta-learners was used for classifying PCOS and NON-PCOS with better accuracy and less time complexity. from that base learner model VGG-16 and from meta learner model XGBoost model as image classifier gives the highest accuracy of classification with is 99. 89%. Four different techniques are used for extracting the features in which the first technique is the traditional approach of ML training which applies relevant digital image processing, the second is chi-square and PCA, and the third technique is using the DNN algorithm and stacking ensemble training. Out of which the DNN algorithm gives better results. Performance analysis was done on the basis of accuracy, precision, recall, and F1 score.

A.K.M. Salman Hosain (2022) et al [2], calculated the accuracy of PCONet- a CNN model to classify polycystic ovarian ultrasound images and then compared it to pre-trained Incep- tionV3.
The same training set and image preprocessing procedure was followed for both models. Both models were trained for thirty epochs. Steps per epoch were determined by dividing the number of images by the batch size. The batch size for training both models was 16. The PCONet showcased

an accuracy of 98.12%, which was higher than the fine-tuned InceptionV3, which showed 96.56% accuracy.

Shazia Nasim and Younas (2022) et al [3], have suggested a technique for PCOS detection using a novel feature selection approach based on the chi-squared(CH-PCOS) mechanism.

Using this approach, the gaussian naive Bayes (GNB) outperformed the ML model and state-of-art-studies. The GNB achieved 100% accuracy, precision, recall, and f1-score with a minimum time complexity of 0.002 seconds. The K-Fold cross-validation of GNB achieved a 100% accuracy score. The study says that the GNB model gives accurate results for the classification of PCOS on the basis of dataset features prolactin(PRL), blood pressure systolic, blood pressure diastolic, thyroid stimulating hormones, relative risk, and pregnancy are the prominent factor having high involvement in PCOS prediction to validate the overfitting of employed ML models they had applied the k-folds cross-validation techniques, the 10 folds of the dataset are used for validation. The technique shows that k-folds validation achieved 100% accuracy and the MLP models achieved 99% accuracy and by k-fold 98% accuracy had achieved. The SGD and KNC models achieved the lowest accuracy.

Subha R (2022) et al [4], have used swarm intelligence (SI) for feature selection and machine learning to develop a robust and efficient diagnostic model to detect PCOS conditions.

The authors have used various methods like correlation and the Chi-Square test for optimal feature selection. They claim, having done a comparative analysis of the results and validation have done based on the parameters accuracy of training and testing, precision, recall, F1-score, and AUC-ROC. They conclude that the feature ML models with different feature selection algorithms are the best for different feature dimensions and the model with PSO-based feature selection gives the highest performance with minimum feature size.

Angela Zigarelli (2022), et al [5], have used Machine learning and deep learning techniques to analyze health data and improve diagnostic accuracy and precision, disease treatment, and prevention.

The goal of their proposed study is to develop a machine-aided self-diagnostic tool that predicts the diagnosis of PCOS with and without any invasive measures, using Principal Component Analysis (PCA), k-means clustering algorithm, and CatBoost classifier. The work is well aligned with

emerging artificial intelligence and digital health care. They claim to have achieved 81% to 82.5% prediction accuracy of PCOS status without any invasive measures in the patient models and achieved 87.5% to 90.1% prediction accuracy using both non-invasive and invasive predictor variables in the provider models. Their proposed prediction models are ultimately expected to serve as a convenient digital platform with which users can acquire pre- or self-diagnosis and counsel for the risk of PCOS, with or without obtaining medical test results. Their model may enable women to conveniently access the platform at home without delay before they seek further medical care. Clinical providers can also use this proposed prediction tool to help diagnose PCOS in women.

Kinjal Raut (2022) et al [6], detected PCOS using various Machine Learning Algorithms like Random Forest, Decision Tree, Support Vector Classifier(SVC), Logistic Regression, K- Nearest Neighbour(KNN), XGBoost with Random Forest (XGBRF), CatBoost Classifier, and Cross-Validation.
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, the conclusion is that CatBoost-Classifier has outperformed and obtained the highest accuracy. A DCNN algorithm with python programming can be a good option for easy identification of PCOS at an earlier stage.

Shubham Bhosale (2022) et al [7], used the DCNN algorithm for detecting PCOS on the basis of ultrasound images. Before applying the CNN algorithm, the data was preprocessed with image segmentation, which is used for reducing the image's noise.
The univariate feature selection method was used for selecting the most suitable features. The time complexity of this algorithm is O(nˆn). The above mathematical model is NP-Complete. The space complexity depends on the presentation and visualization of discovered patterns. More the storage of data, the more the space complexity.

The proposed method by Wenqi Lv and Huang. (2022) et al [8], is composed of image preprocessing, feature extraction, and classification steps based on deep learning.
Used an improved U-Net embedded with an attention module to segment the sclera from full-eye images, a Resnet18 to extract deep features, and a multi-instance learning model to classify PCOS and normal samples are made. Results show that the non-invasive screening method achieved a mean AUC of 98%, a mean accuracy of a dataset that contains 721 subjects.

The paper presented by Bhat (2022) et al [9], describes the various methods used in medical image preprocessing for the detection of PCOS.

No of follicles, their size, shape, and properties are important factors that can be grasped from ultrasound images. The ultra-sound images can be noisy and can degrade the image quality. The authors analyzed the image enhancement methods such as Histogram equalization, Adaptive histogram equalization, Contrast stretching. Histogram equalization is based on the principle that for a better image, its histogram will be normally distributed. This method is well-suited for grayscale images. Adaptive histogram equalization is an enhanced version of normal HE. HE works with a single histogram of an image whereas it divides the image into several parts and constructs the histogram. In contrast stretching, the contrast values are fitted into the desired range. The authors also discussed Particle swarm optimization and hybrid principle component analysis for better classification.

Kodipalli and Devi (2021) et al [10], studied women's health in women under the age of 25.

The majority of research was carried out using traditional methods like the T-square test, and the Chi-Square test. The implementation used is an approach to apply machine learning and fuzzy systems logic to the data and perform a comparative study of the two. The data collection for this study included both physical assessments (menstrual cycle, regularity of the cycle, length of the cycle, duration of the cycle, recent weight gain, hair loss, family history of having diabetes and hypertension, and eating and sleeping habits) and psychological assessments (anxiety, depression, body image dissatisfaction). For this research work the comparison matrix Mij is constructed, then the fuzzified geometric mean value is calculated, Calculate the fuzzy weights, Any defuzzification method can be used to calculate the defuzzified weights, From the weights (wi), calculate normalized weights. Then these weights (wi), calculate normalized weights. Naive bias, Decision tree, and Random forest have an accuracy of 97.65%, 96.27%, and 89.02% respectively. The presented study proved that 66.07% of women with PCOS have associated mental health issues.

In a comparative study of different denoising techniques by Shruti Bhargava Choubey (2021) et al [11], noises that are used for evaluation are standard as they give their mere presence in almost every case of imaging.

The study of dual-stage filtering images for the medical field has been executed efficiently. The MSE, PNSR, and WPSNR were used to evaluate the system. The noise was analyzed with a focus

on its ill effects in PCOS Images that can lead to the evaluation of diseases. The PCOS test images had some development in most of the parameters in deliberation.

Vikas B (2021) et al [12], followed the iterative process.
In every iteration, accuracy is compared with the previous iteration. The models used are the basic CNN (benchmark) model, a model with more hidden layers and dropout layer, transfer learning model trained on the augmented dataset with hyperparameter tuning. The accuracy significantly improved by 10% from the initial model. The highest accuracy obtained was 94%.

VGG-19, DenseNet-121, ResNet-50, and inception V3 and model stacking, the GAN(Generative Adversarial Network) architecture is used by Kumari (2020) et al [13], to produce artificial images for better performance.
The model was with VGG-19, DenseNet121, ResNet-50, and inception V3 and model stacking, out of this highest accuracy with better sensitivity and specificity is achieved by VGG-19 i.e. approximately 70%. Due to less amount of dataset, a technique of synthetic image generator along with the data augmentation is used

Tanwani [14] (2020) compared machine learning algorithms K-Nearest Neighbor (K-NN) and Logistic Regression.
The method to find accuracy was an F1 score for both of the algorithms. The F1 score helped determine the best model between the two. The F1 score for KNN was found to be 0.90 and for that of Logistic Regression is 0.92, Therefore, the model of Logistic Regression was selected for the diagnosis of polycystic ovary syndrome detection in ovaries.

Holger H. (2018) et al [15], presented an up-to-date overview of semi-supervised learning concepts considering earlier and recent advancements in machine learning.
Semi-supervised learning attempts to improve the performance of models. This technique has been proven to be best for computer-aided disease detection, part of speech tagging, and drug discovery. The authors have discussed assumptions about the data necessary for being able to apply the SSL. Further, the taxonomy and different methods of SSL are described with analogies. Pseudo-labeling is one of the algorithms that first train the model using a labeled dataset. The model with better accuracy is used to label the unlabeled dataset. The model is again trained on pseudo-labeled data to attain higher performance.

2.2 Literature Review Summary

Sr.No.	Title of paper	Year	Authors	Findings	Technologies
1.	A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics	2022	Shazia Nasim Mubarak Almutairi Kashif Munir Ali Raza Faizan Younas	This paper explains the technique for PCOS detection using a novel feature selection approach based on chi-squared(CH-PCOS) mechanism. The K-fold methods performs with the highest accuracy, followed by MLP models.	Gaussian naive bayes (GNB), K-Fold cross-validation, MLP models
2.	Deep Learning Algorithm for Automated Detection of Polycystic Ovary Syndrome Using Scleral Images	2022	Wenqi Lv, Ying Song , Rongxin Fu , Xue Lin , Ya Su, Xiangyu Jin , Han Yang, Xiaohui Shan , Wenli Du, Qin Huang , Hao Zhong , Kai Jiang , Zhi Zhang, Lina Wang and Guoliang Huang.	This paper explains the use of an improved U-Net embedded with an attention module to segment the sclera from full-eye images, a Resnet18 to extract deep features, and a multi-instance learning model to classify PCOS and normal samples are made. Results show that the non-invasive screening method achieved a mean AUC of 98%, a mean accuracy of a dataset that contains 721 subjects.	Resnet18 model

Sr.No.	Title of paper	Year	Authors	Findings	Technologies
3.	PCOS Detect using Machine Learning Algorithms	2022	Kinjal Raut, Chaitrali Katkar, Prof. Dr. Mrs. Suhasini A. Itkar.	This paper includes use of various Machine Learning algorithms like Random Forest, Decision Tree, Support Vector Classifier(SVC), Logistic Regression, K-Nearest Neighbour(KNN), XGBoost with Random Forest (XGBRF), CatBoost Classifier, and Cross-Validation. The accuracies obtained by different algorithms are: Decision Tree, SVC, Random Forest, Logistic Regression, K-Nearest Neighbors, XBRF, CatBoostClassifier. Therefore, from the above results, the conclusion is that CatBoost-Classifier has outper-formed and obtained the highest accuracy. A DCNN algorithm with python programming can be a good option for easy identification of PCOS at an earlier stage.	Machine Learning Algorithms - Random Forest, Decision Tree, SVC, Logistic Regression

Sr.No.	Title of paper	Year	Authors	Findings	Technologies
4.	PCOS (POLYCYSTIC OVARIAN SYNDROME) Detection Using Deep Learning	2022	Shubham Bhosale, Lalit Joshi, Arun Shivsharan	In this paper application of the CNN algorithm practiced with the data, which was preprocessed with image segmentation, which is used for reducing the image's noise. The univariate feature selection method was used for selecting the most suitable features. The time complexity of this algorithm is O(nˆn). The above mathemat- ical model is NP-Complete. The space complexity depends on the presentation and visualization of discovered patterns. More the storage of data, the more the space complexity.	CNN algorithm
5.	Polycystic Ovarian Syndrome Detection by Using Two-Stage Image Denoising	2021	Shruti Bhargava Choubey1, Abhishek Choubey1, Durgesh Nandan2*, Anurag Mahajan	The study of dual-stage filtering images for the medical field has been executed efficiently. The MSE, PNSR, and WPSNR were used to evaluate the system. The noise was analyzed with a focus on its ill effects in PCOS Images that can lead to the evaluation of diseases. The PCOS test images had some development in most of the parameters in deliberation.	MSE, PNSR and WPSNR evaluation methods

Table 2.1: Literature survey

Chapter 3

SOFTWARE REQUIREMENTS AND SPECIFICATIONS

3.1 Introduction

3.1.1 Project Scope

Detection of PCOS involves a number of factors and tests such as blood test, genetic test, obesity etc. One of such tests is ultrasound test also known as sonography. This project is based on the images obtained by ultrasound test. In it we aim to consider more tests. Along with that the PCOS has several types. The scope of this project is limited to detection of PCOS in females. In future we will try to build a system for detailed analysis of PCOS like which type it is, what stage, what are its side effects etc. As sometimes females hesitate to go to the doctor for their problems so it is helpful for females to know whether they have PCOS or not by taking small quizzes or if they have sonography images by entering that they are able to detect the disease.

3.1.2 User Class and Characteristics

The user will have the certain characteristics for handling the system that the user should be a female or a doctor as it is a system which is able to detect the PCOS in females only because of it only females are able are the user of the system and along with that the doctor will also be able to handle the system for checking that the patients have a disease. As the detection of PCOS is quite difficult and it takes time for detection of the PCOS so the doctor will be able to identify the patient will have PCOS by entering the sonography report image on a system within the less time.

3.1.3 Assumptions and Dependencies

The assumption made by us is that the user should enter the appropriate image in the system. It should be only an ovary image or a sonography image of the user only then our system will be able to accurately classify the PCOS otherwise it will give the wrong results . If the user will not have the image then by using other functionality of taking quiz users will be able to get the correct result only by entering the correct responses for those questions ,if they are giving inaccurate responses then the system will not be able to identify the PCOS . Based on accuracy of user input, our system's other functionalities will work if the user enters the correct input then only our build system will be able to give the correct results.

3.2 Functional Requirements

3.2.1 System Features

The Functional requirement for the first module of the system that the user will be able to enter the input image successfully. If entering the image system fails then it creates a problem as based on only the entered input the user will be able to get the result. If the system becomes inefficient for accepting the input then there is no use for t6he build system. Similarly if the user will choose the option of detecting PCOS by taking the quiz then the user will be able to take the quiz, if it fails and the user will not be able to take a quiz then the system is inefficient in building . The conclusion point for the functional requirement of the system is that users will be able to enter the image successfully and they will be able to take the quiz successfully then only the system will work and give the correct results.

3.2.2 Communication Interface

The system will have the user-friendly interface so that the user handles the system functionalities smoothly. The user interface of the system contains a small video as a user guide which will give users a basic idea how that system will exactly work?, what steps users need to follow to get the results. The communication interface will show users steps to follow along with it and provide the basic information about the PCOS , what are the different kinds of solutions on it . It provides two options for users for knowing about the results. The first is that it gives the option for uploading the image of an ovary or a sonography image if the user will not have the image then it provides another option to the user to take a small quiz. If the user will have the PCOS then the system will show the user the nearest doctors to them so that they can contact them and get

the appropriate treatment on it.

3.3 Non-Functional Requirements

The website will be advanced and easy to use. Each page will load within 2 seconds. According to the requirement of result and input format, i.e. text or image the user will be provided with the interface. All the fields will be mandatory and after clicking on the button of submit, the result page will be displayed within 2-3 Seconds. The website will be device friendly and does not have other specific expectations of the environment.

3.3.1 Performance Requirements

The platform used for the application will reduce the response time. As for the feature of diagnosis through image, the image should be of jpg, png, jpeg format. The Sonographic image must be clear.

3.3.2 Safety and Privacy Requirements

The application will ask for some personal hygiene and menstrual cycle related questions. The information provided by the user will remain protected. Even in our feature where the diagnosis is done by Sonographic images, those uploaded images will not be stored for users' privacy. The whole process will be secure and the user will remain anonymous.

3.3.3 Security Requirements

On starting a session, users will not have to sign in to use the website. As the user doesn't have to provide the personal information, security will be unharmed. The website will not require any third party applications.

3.3.4 Software Quality Attributes

Even for multiple sessions of application on different systems, the application will provide high end results in 2-3 seconds. The application will not be out of reach even when the system is getting updated in the backend. The updated version of the application will be automatically updated so that users will not have to face obstacles. The application is environment independent and it gives accurate results 99.

3.4 System Requirements

The initial stage of model training and model testing is performed on kaggle platform. The application will mainly be created in the python language using flask framework.

3.5 Agile Methodology

We have studied and performed the model with the process of planning, Analysis, Designing, Building, testing. We have been using an Agile model. As the agile model promotes teamwork and Functionality can be developed rapidly and tested frequently. The Agile model is based on the adaptive software development methods. There is feature driven development and we have adapted to the changing product requirements dynamically. The product is tested very frequently to minimize the risk of any major failures in future.

3.6 System Implementation Plan

1. The research was done with the help of an already published research paper on similar topics.

2. The data has been collected with the help of google from the women of our college. The data is validated from the doctor.

3. Compare models of machine learning and deep learning for high accuracy, choose one model to work with.

4. Training and testing of models for both textual input data and image input data.

5. Make a frontend for an application which is easy to use. Integrate the machine learning and deep learning models.

6. Test the application against all kind of data.

7. Deploy the application.

Chapter 4

SYSTEM DESIGN

4.1 System Architecture

The below diagram gives the system architecture for PCOs detection systems. In which the User will visit the website or the GUI that we are making for creating a User-friendly environment. There are two types of input options provided to the user one is the ultrasound images of sonography and if the user doesn't have the ultrasound image then the simple quiz provided to the user in which they give the answer of general 32 questions like are they have their regular periods? What is the age? What do they describe about their body?, etc, Firstly the input from the user is taken according to the type of input preprocessing is done on the input f the input is the ultrasound images then the preprocessing is done on that images such as reshaping the image, rescaling and other preprocessing steps are performed to processes the preprocessing of the image. Then segmentation is used for by which the image is divided into multiple parts or segments and for each segment a separate label is provided which acts as the feature for that segment then using different feature extraction methods high-level features are extracted from the image which is used to identify the PCOS and Non-Pcos images. The CNN model is then used to train the model with the different kinds of transfer learning algorithms. Based on the model performance any one of the models is selected to train the images. If the input is not in image format then the simple questions are asked to the user and the response from the user is stored. Then feature extraction and analysis is performed to identify the important features for classifying their response are PCOS or Non-PCOS. The most important features are extracted. Depending on the response for the dataset a suitable pre-trained machine learning model is used such as a supervised, unsupervised, or semi-supervised algorithm based on the labeling of the data.

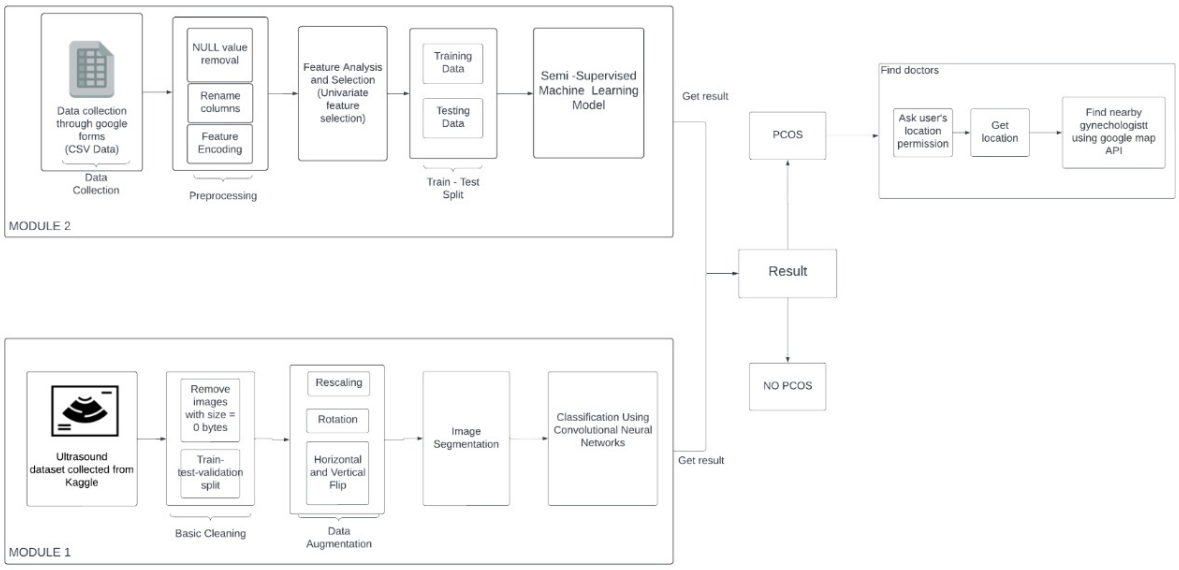


Figure 4.1: System Architecture

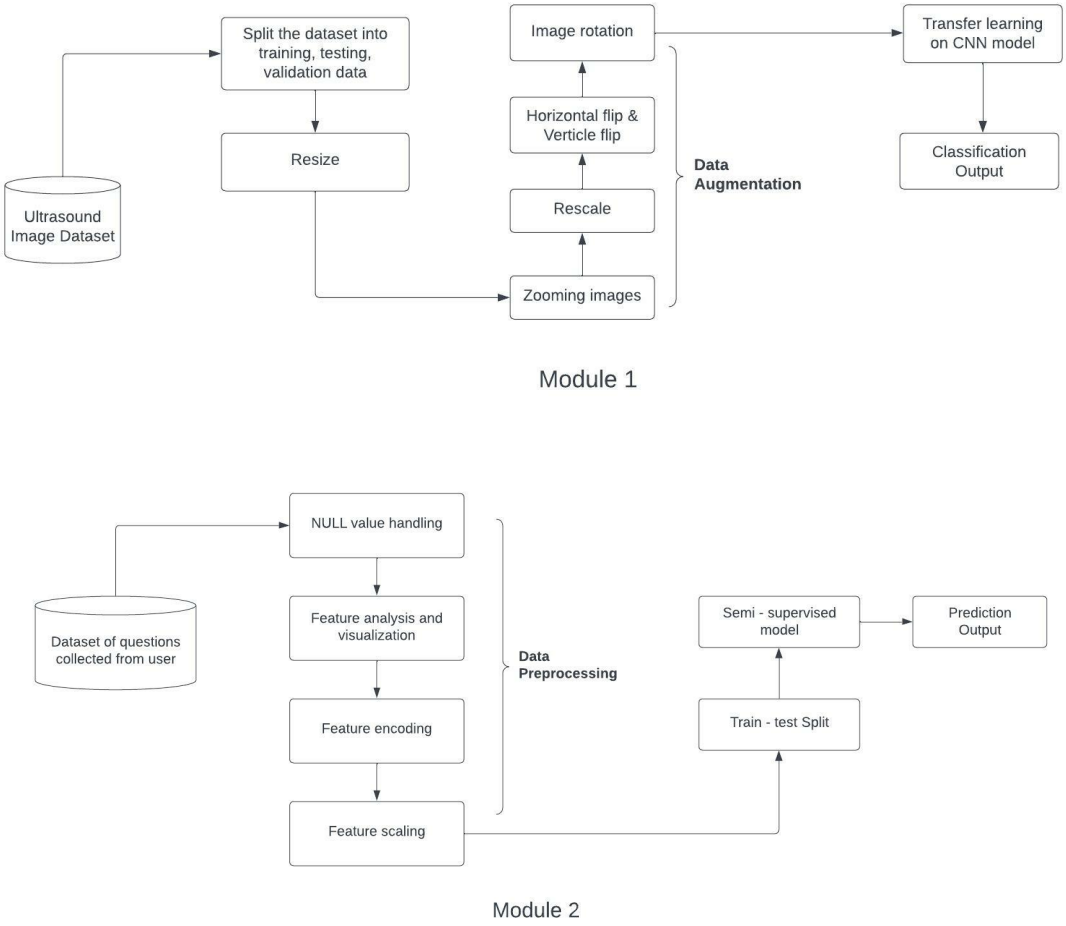


Figure 4.2: Data Flow Diagram

4.2 State Transition Diagram

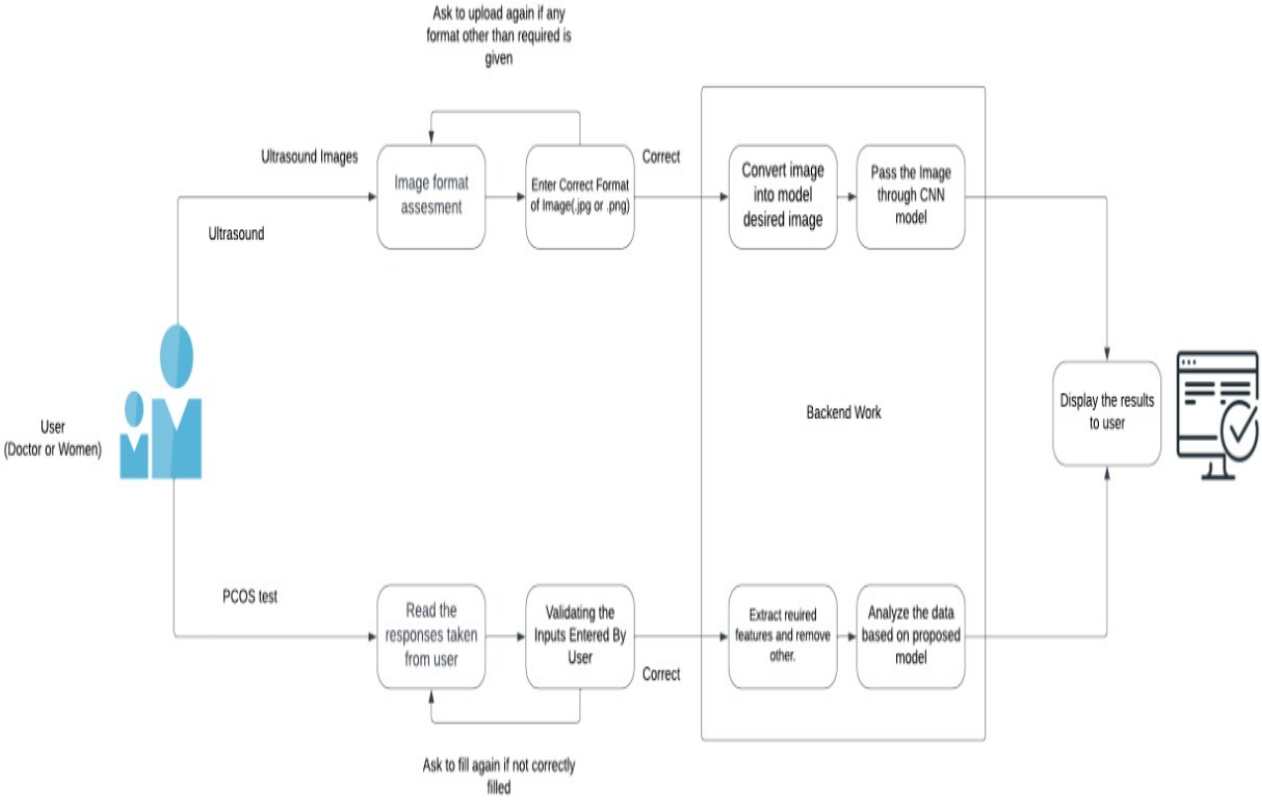


Figure 4.3: State Transition Diagram

Chapter 5

PROJECT PLAN

5.1 Project Estimate

5.1.1 Cost

The cost of developing a PCOS prediction system would depend on various factors such as the size and complexity of the system, the technology stack used, the number of features and algorithms involved, and the level of accuracy required.

Cost of data:

The cost of data collection and preparation can be a significant factor in developing a PCOS prediction system. This can include the cost of acquiring and cleaning medical data from various sources, such as electronic health records, laboratory results, and patient surveys.

Cost of research:

This can include the cost of hiring researchers and developers, software and hardware costs, and other expenses associated with the development process.

Cost of production:

If the system is deployed on-premise, there will be costs associated with maintaining hardware and software infrastructure. If the system is deployed in the cloud, there will be ongoing costs for cloud hosting services and any additional services required to maintain and scale the system.

5.1.2 Time

Including the complexity of the system, the amount of data to be processed, the number of features and algorithms involved, and the development team's size and experience.
The development time for a PCOS prediction system could range from several months to a year or more. This timeline includes the following stages:

Sr.No.	Stages	Estimated Time
1.	Problem understanding and requirements analysis	2 Months
2.	Data collection and preparation	2 Months
3.	Feature engineering and algorithm selection	2 Months
4.	Model training and testing	2 Months
5.	Deployment and integration	2 Months
	Total estimated time	8 Months

Table 5.1: Time

5.1.3 Size and Scope :

Tasks to be done:

- Data collection
- Data exploration
- Model building
- Front-end development
- Integrating the model with frontend
- Documentation

Required expertise:

- Data Scientist
- Data Analyst
- ML Engineer
- Frontend Developer
- Doctor mentor

The scope of the project is well understood by team members and requirements are clearly mentioned.

5.2 Project Resources

- **Stakeholders:** 4 Team Members, 1 Internal Guide
- **Development Libraries:** Tensorflow, Keras, Sklearn, OpenCV
- **Development Platforms:** Python Notebook, Visual Studio Code
- **Technologies:** Machine Learning, Deep Learning, Flask, HTML, CSS, JavaScript

5.3 Risk Management

5.3.1 Risk Identification

Identification of risk is an important parameter for the project completion. Risk can be affect the overall performance of the system, it may damage the system and by the end user will not be able to use the build system. For the effective use of the system by the user we should need to consider all kind of the risk. Their may be technical,cost, schedule or any other kind of risk.

1. **Technical Risk:** It is a functional risk which impact the performance of the system . If the data is biased which menace dataset contain more data of particular class of PCOS such as more no. of images of PCOS as compared to NON-PCOS images, then their might be a possibility that result will always show PCOS class for NON-PCOS image too.

2. **Cost Risk:** It is kind of risk which aries due to less funds are generated for the project . The cost risk for the project is the cost required for handling the GPU resources. It also contain the cost risk as cost required for the collection of data from the pathology lab need to pay .

3. **Schedule Risk:** It is kind of risk which may be arises by not following the deadline for the completion of task associated with the project . It contain learning new prerequisite such as flask may take a time and which may cause delay in project completion. The another risk associated with the project is collection of data may take a long time which again impact on the completion of the project .

5.3.2 Risk Analysis

It contains the identification of the risk impact onto the project. By considering the above risk analysis of that risk is important to reduce the exposure and minimize the impact. As all the technical risk mentioned above have low impact on the project if the dataset contains the balance data for both the PCOS and NON-PCOS class. The impact of cost risk associated with utilizing the GPU resources is moderate and for collecting the data from the pathology lab has high impact on the project. For the schedule risk it has high impact on the project by considering the appropriate deadline for the task completion may reduce the impact of the risk.

5.3.3 Overview of Risk Mitigation, Monitoring, Management

By considering the above identified risk associated with the project if the data is collected manually or utilizing the standard available dataset from the kaggle or any other website instead of collecting the data from pathology lab may not cause the cost risk of collecting data from pathology lab. By collecting the data for other class of PCOS may reduce the technical risk it is done by collecting the more data for balancing both the classes of the PCOS and NON-PCOS. The solution for the schedule risk is that instead of learning flask prerequisite by all the team members only some of the team member may learn itt and other may focus onto the data collection process which will handle the schedule risk and reduces the impact of these risk onto the project. The risk management and monitoring is an important part by collecting frequent meeting for the risk monitoring reduces it's impact and identify the another solution for handling the risk. Frequent meetings for the following the appropriate deadline for competition of rtask may also result in reducing the risk associated with the project.

5.3.4 Project Task Set

The distribution of effort across the software process was 40% allocated to analysis design, 20% allocated to coding, 40% allocated to testing.
The PCwomenOS project is mixture of Concept Development Project and New Application Development Projects, the project is undertaken as a consequence of the development of new idea and application of technology.
The project was approached by applying the following major tasks:

1. **Concept Scoping:** As per the initial stage, we researched about the concept. We studied the scope of the concept and other areas related to it. In the concept scoping stage, we read a lot of research papers related to women health issues and the ways we can use data science concepts to solve women menstrual health issues.

2. **Preliminary concept planning:** The scope of using data science into solving women menstrual health issues is wide. So we focused on polycystic ovarian syndrome menstrual diagnosis using data science.

3. **Technology and risk assessment:** The technology decided for the PCOS diagnosis in Machine learning and Deep learning. The risks were analyzed.

4. Proof of concept: The data which was collected for machine learning and deep learning was verified by the gynecologist. The machine learning and deep learning algorithms were studied for the implementation of the concept.

5. **Concept implementation:** The concept of using machine learning model for symptoms based pcos prediction and deep learning model for ultrasound image based pcos prediction was done. Additional features were implemented. The application was tested against several inputs and the results were accurate.

6. **Customer Reaction:** After the testing, the application was made available to the user and the expected results matched to the results generated by the application.

5.4 Project Schedule

5.4.1 Task Network

The flow of the project started with deciding on the concept of Polycystic ovarian syndrome diagnosis using data science. We referred to research papers. We collected the data for the machine learning model by circulating google form in our college. Later, we started working on implementation. We worked on module 1 and module 2 in parallel. The application was tested against multiple inputs and the application was made available after the deployment.

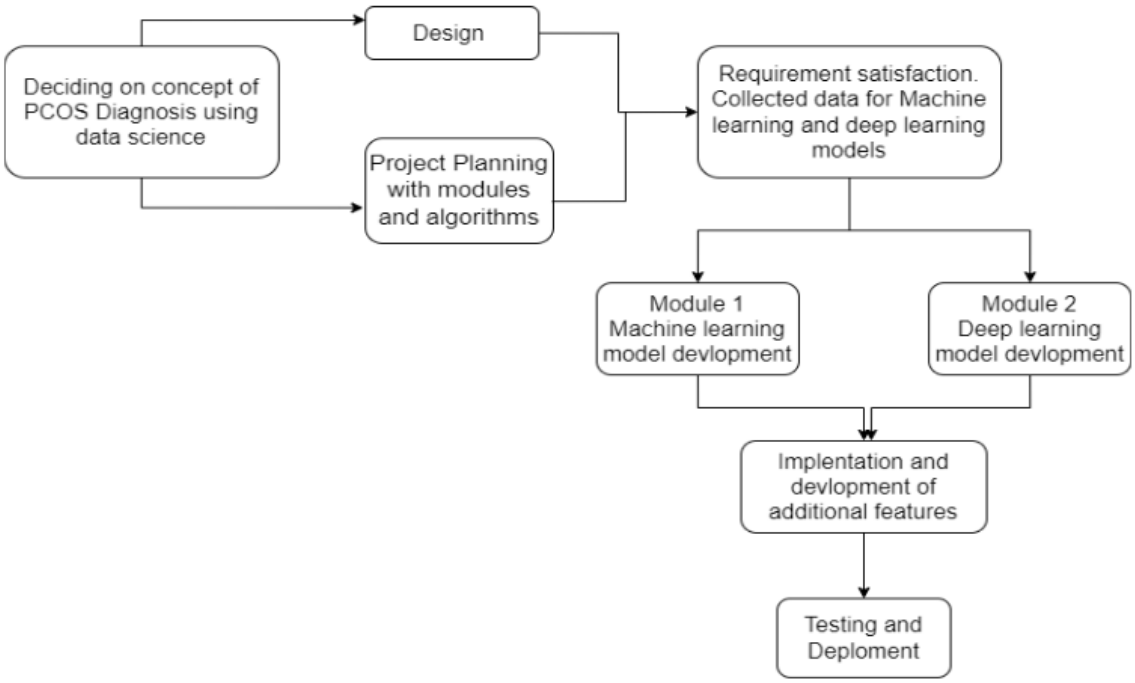


Figure 5.1: Task Network Diagram

5.4.2 Timeline Chart

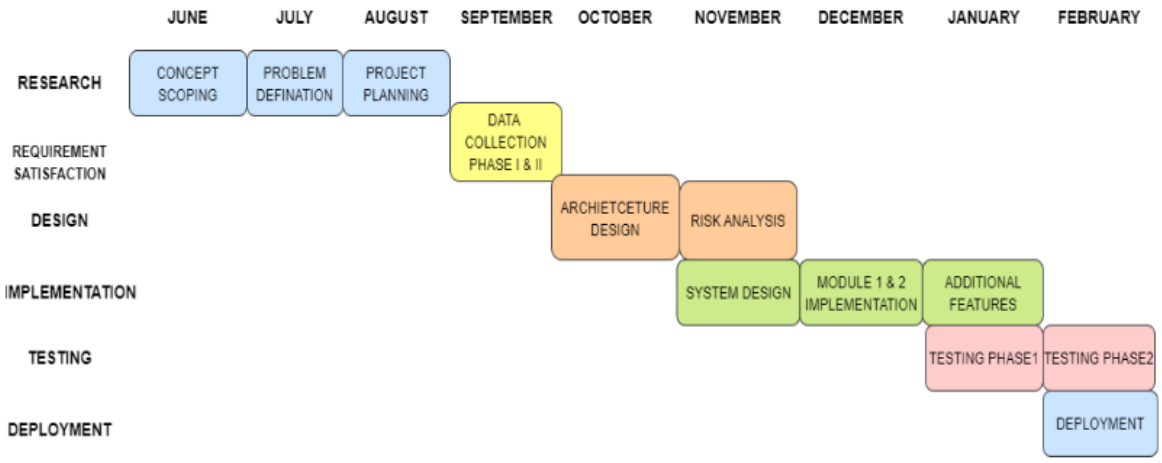


Figure 5.2: Timeline Chart

5.5 Team Organization

5.5.1 Team structure

Our team structure is democratic team where input from each member is taken for a significant decision. Group leadership revolves only among the group members. The structure allows input from all members of the group, which can lead to better decisions in various problems.

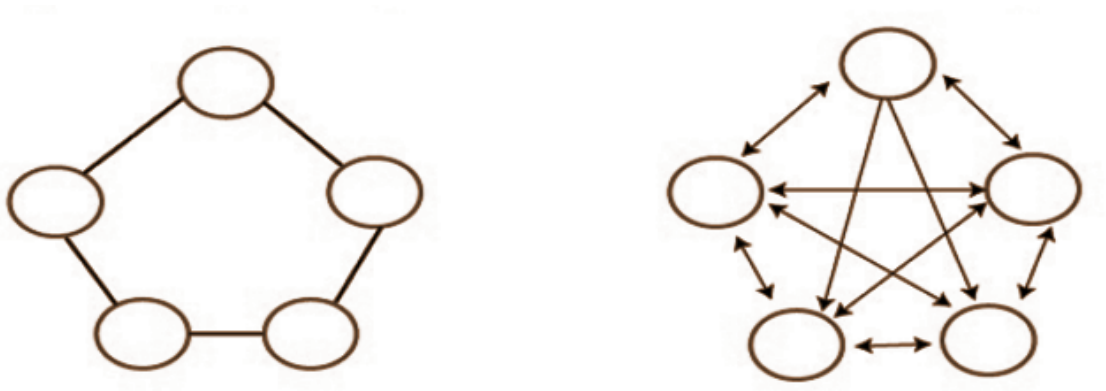


Figure 5.3: Team Structure

The left diagram represents the structure of democratic team and the right diagram represents the communication path.

5.5.2 Management reporting and communication

The management reporting and communication was done on the basis of following report

- **Team availability report:** The report was updated by every member according to the availability of that member. According to this report the meetings were scheduled.
- **Status report:** The status report is updated after every task completion by that team member. The subtask is also updated in this report.
- **Project health report:** This presented the execution status of the project.
- **Time tracking report:** It shows the time spent on a task by team members within project development.

Chapter 6

PROJECT IMPLEMENTATION

6.1 Overview of Project Modules

There are major 2 modules in this project:

1. Deep learning-based PCOS detection using ultrasound images This module allows uploading an image of an ultrasound report in jpg/png format. Deep learning-based disease detection has been proven to be the best in the medical sector. Hence this module uses neural networks to detect PCOS. The result is displayed to the user in an understandable form without going into medical terminologies.

2. Machine learning-based PCOS self-test This module allows users to simply take a test that predicts the likelihood of PCOS. The test consists of questions related to the symptoms faced, lifestyle habits, and previous medical conditions. Each question has a drop-down answer option. This module has been designed using a non-traditional algorithmic category of machine learning called semi-supervised learning.

6.2 Tools and Technologies Used

- **OS:** Windows The system is designed on the windows operating system with configuration.

Windows Edition	11
Version	21H2
Installed RAM	8GB
System type	64-bit operating system
Processor	Intel(R) Core(TM) i5-1035G1 CPU

Table 6.1: Operating System

- **Google Collaborative Notebook:** Data exploration, visualization, analysis, and preprocessing were conducted on google python notebooks. Further model selection and model building were also done using google notebook. Google notebook was selected because of its ability to run the code cells individually and its collaborative nature.

- **Visual studio code:** Visual studio code is an integrated development environment that supports multiple languages and has a powerful debugging environment. VS code is used to design the front end of the system as well as to connect the model with the front end. VS code was selected because of its features such as extensions, live server mode, etc.

- **Python:** Python is a programming language that has applications in various domains. It is the most used language for data science applications. Python has various libraries and modules such as pandas, matplotlib, numpy, sklearn, etc. Along with that, Flask is a framework of python which is used for developing the front end of the system.

- **HTML, CSS, JavaScript:** HTML is used for developing the website. CSS along with bootstrap is used to make the site responsive and attractive. Javascript is used for showing the nearby gynecologist.

6.3 Dataset Description

There are major 2 modules in this project:

1. Deep learning-based PCOS detection using ultrasound images The dataset of ultrasound images available on Kaggle, which includes two classes of images: Infected and Non-Infected. The infected class has 1,562 images and the non-infected class has 2,284 images, each divided into training, testing, and validation sets. The images show a visible difference in their structure, with infected images showing a net-like structure that represents cysts in ovaries or follicles, while non-infected images have fewer follicles or lack the net-like structure. This is an important distinction as it is a significant indicator of Polycystic Ovary Syndrome (PCOS), a hormonal disorder that affects women.

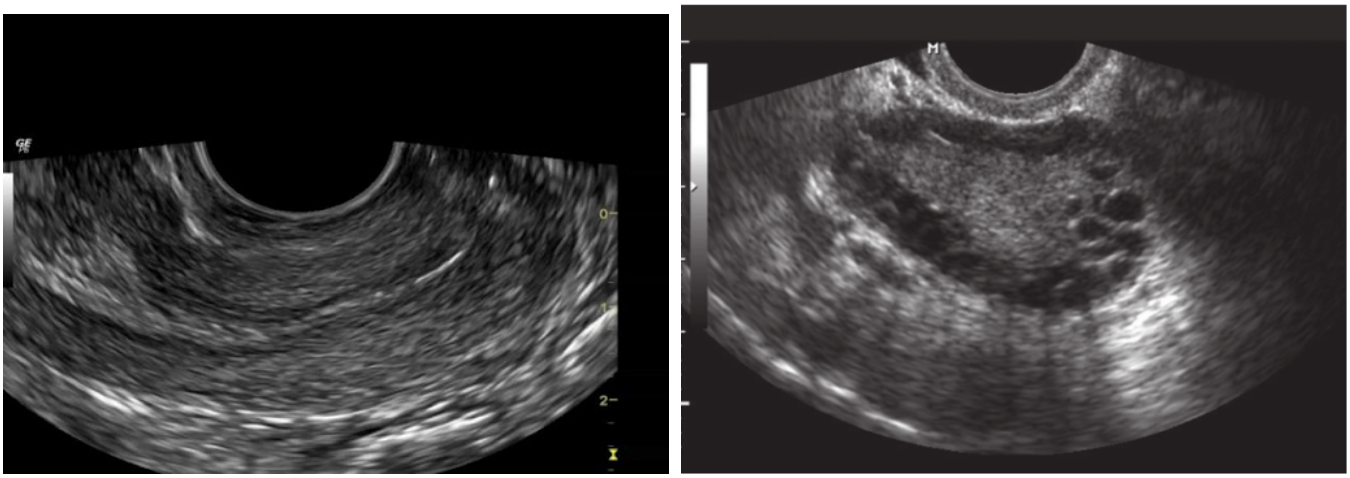
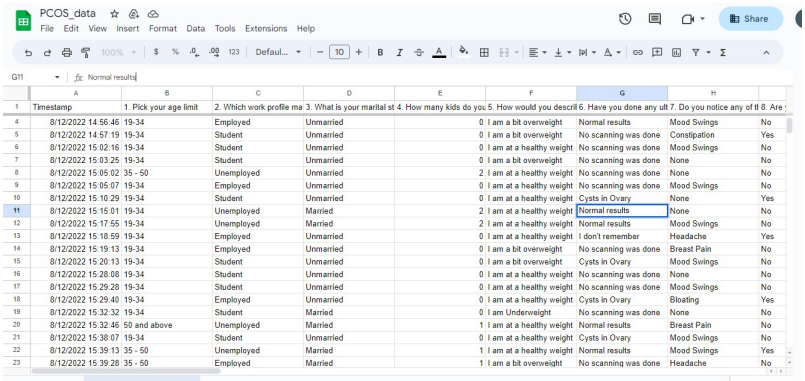


Figure 6.1: Ultrasound Image Data

2. Machine learning-based PCOS self-test Records from the dataset are collected by floating the Google form. The form has 167 responses till now. The form consists of 23 questions with answers that result in various features such as age, marital status, body description, number of kids, menstrual cycle duration, and pain during the menstrual cycle. Based on the features of the previous responses, the new response is classified into either PCOS or Non-PCOS.



	Timestamp	1. Pick your age limit	2. Which work profile ma	3. What is your marital st	4. How many kids do you	5. How would you descril	6. Have you done any ult	7. Do you notice any of tl	8. Are
4	8/12/2022 14:56:46	19-34	Employed	Unmarried	0	I am a bit overweight	Normal results	Mood Swings	No
5	8/12/2022 14:57:19	19-34	Student	Unmarried	0	I am a bit overweight	No scanning was done	Constipation	Yes
6	8/12/2022 15:02:16	19-34	Student	Unmarried	0	I am at a healthy weight	No scanning was done	Mood Swings	No
7	8/12/2022 15:03:25	19-34	Student	Unmarried	0	I am a bit overweight	No scanning was done	None	No
8	8/12/2022 15:05:02	35 - 50	Unemployed	Unmarried	2	I am at a healthy weight	No scanning was done	None	No
9	8/12/2022 15:05:07	19-34	Employed	Unmarried	0	I am at a healthy weight	No scanning was done	Mood Swings	No
10	8/12/2022 15:10:29	19-34	Student	Unmarried	0	I am at a healthy weight	Cysts in Ovary	None	Yes
11	8/12/2022 15:15:01	19-34	Unemployed	Married	2	I am at a healthy weight	Normal results	None	No
12	8/12/2022 15:17:55	19-34	Unemployed	Married	2	I am at a healthy weight	Normal results	Mood Swings	No
13	8/12/2022 15:18:59	19-34	Employed	Unmarried	0	I am at a healthy weight	I don't remember	Headache	Yes
14	8/12/2022 15:19:13	19-34	Employed	Unmarried	0	I am a bit overweight	No scanning was done	Breast Pain	No
15	8/12/2022 15:20:13	19-34	Student	Unmarried	0	I am a bit overweight	Cysts in Ovary	Mood Swings	No
16	8/12/2022 15:28:08	19-34	Student	Unmarried	0	I am at a healthy weight	No scanning was done	None	No
17	8/12/2022 15:29:28	19-34	Student	Unmarried	0	I am at a healthy weight	No scanning was done	Mood Swings	No
18	8/12/2022 15:29:40	19-34	Employed	Unmarried	0	I am at a healthy weight	Cysts in Ovary	Bloating	Yes
19	8/12/2022 15:32:32	19-34	Student	Married	0	I am Underweight	No scanning was done	None	No
20	8/12/2022 15:32:46	50 and above	Unemployed	Married	1	I am at a healthy weight	Normal results	Breast Pain	No
21	8/12/2022 15:38:07	19-34	Student	Unmarried	0	I am at a healthy weight	Cysts in Ovary	Mood Swings	No
22	8/12/2022 15:39:13	35 - 50	Unemployed	Married	1	I am at a healthy weight	Normal results	Mood Swings	Yes
23	8/12/2022 15:39:28	35 - 50	Employed	Married	1	I am a bit overweight	No scanning was done	Headache	No

Figure 6.2: ML Module Data

6.4 Algorithm Details

6.4.1 Semi-Supervised Learning

Semi-Supervised learning is a category of machine learning algorithms. In semi-supervised learning, the dataset is a mixture of labeled and unlabeled data. The model is first trained on the labeled dataset using supervised learning algorithms. The model with the best accuracy is selected for labeling the unlabeled data. Again the model is trained on the entire combined dataset to achieve better performance. This technique is called pseudo-labeling.

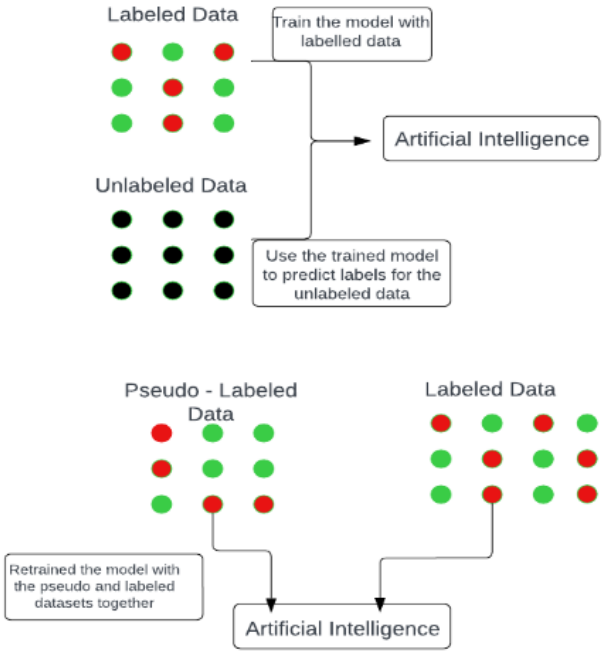


Figure 6.3: Semi-supervised Learning

Algorithms

Logistic Regression : Logistic regression is used for the classification in which the categorical data divides into the labeled class . It is a supervised machine learning algorithm . Logistic regression is used to find the relationship between the dependent and the independent variables. By predicting the independent variables it will assign the label class. Logistic regression model implemented on the dataset which gives the accuracy of 94.12% accuracy by using the above mentioned approach of semi supervised machine learning.

KNN: KNN stands for the k- nearest neighbors It is a supervised machine learning algorithm used for the classification of the data into the PCOS and Non-PCOS class. It will calculate the distance between the data point and the centroid and the data point will assign it to the nearest neighbor centroid. Here "k" stands for the number of neighbors. KNN model implemented on the dataset which gives the accuracy of 97.06% accuracy by using the above mentioned approach of semi supervised machine learning.

Decision Tree: It is a supervised machine learning algorithm it is used for the classification of the data into the PCOS and Non-PCOS class.It is a tree like structure which divides the dataset features into a branch and the end node call as a leaf node are the class for that divides features set . each feature is divided into the node and finally different implemented on the dataset which gives the accuracy of 88.24% accuracy by using the above mentioned approach of semi supervised machine learning implemented on the dataset which gives the accuracy of 92.83% accuracy by using the above mentioned approach of semi supervised machine learning.But it is not used for actual classification as it is just used to know on feature dataset get divided into the label class.

Random Forest: It is a supervised machine learning algorithm used for the classification of the data into the PCOS and Non-PCOS class. It is a collection of multiple decision trees .Decision tree is a tree-like structure which divides the dataset features into a branch and the end node called a leaf node is the class that divides features set . each feature is divided into the node and finally different implemented on the dataset which gives the accuracy of 88.24% accuracy by using the above mentioned approach of semi supervised machine learning implemented on the dataset which gives the accuracy of 92.83% accuracy by using the above mentioned approach of semi supervised machine learning.But it is not used for actual classification as it is just used to know on feature dataset get divided into the label class.

Selected Approach

KNN is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It is used to model the probability of a certain event occurring, typically binary outcomes (yes or no, success or failure, etc.), based on the values of the independent variables. The KNN model uses a sigmoidal function to model the relationship between the independent variables and the probability of the binary outcome. In this case, the independent variables would be the various symptoms associated with PCOS, and the dependent variable would be the presence or absence of the condition. Out of 24 independent variables i.e symptoms the top 5 variables/features are used for training the model.

Sr.No.	Features	Correlation
1.	Regular period	0.528616
2.	Period cycle	0.572458
3.	Period pain scale	0.351006
4.	Excess hair growth	0.314092
5.	Stress	0.316228

Table 6.2: Top 5 relevant features

The model is initially trained on 100 data points and tested on 26 data points with an accuracy of 96.15%. This model is used to make predictions on completely unseen data with 40 data points. Further, the model is retrained on a dataset generated by combining initial 100 data points and 40 predicted data points. The overall accuracy obtained is 97.06%.

6.4.2 Deep learning

Deep learning is a category of machine learning. Deep learning is based on neural networks which mimic human brain behavior. Deep learning has different algorithms such as artificial neural networks, recurrent neural networks, generative adversarial networks, etc.

Algorithms

- Convolutional Neural Networks

CNN is a category of neural networks that use convolutional layers, and pooling layers to perform the desired tasks. The potential of CNN is in its convolutional layers which can perform great feature extraction. is a well-suited algorithm for image classification tasks specifically in medical image classification problems. It has been proven to be the best for early detection and diagnosis of various diseases such as cancer, brain tumor, cardiac arrest, etc.

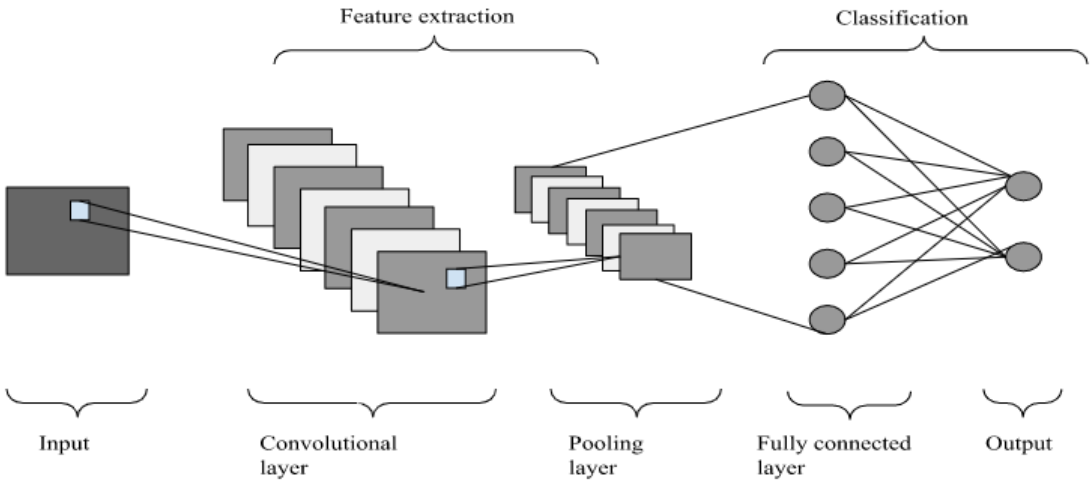


Figure 6.4: Architecture of CNN

VGG16: Visual Geometry 16 : VGG16 is a widely used CNN model for image recognition tasks. It consists of 13 convolutional layers and 3 fully connected layers. The convolutional layers use 3X3 kernel filters and the max pooling layers use 2X2 windows. VGG16 is trained on the ImageNet dataset and achieved a state-of-the-art performance of about 92% accuracy.

RESNET50: Residual Network 50 : RESNET50 is a DCNN model introduced in 2015. This architecture is comprised of 50 layers including convolutional layers, batch normalization layers, activation function, and max-pooling layers. RESNET uses the technique called 'Residual Mapping' which basically measures the difference between the input and output of the layer and thereby achieves significant performance in training deep networks.

INCEPTIONV3 : INCEPTION V3 was introduced as part of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). This architecture uses 48 convolutional layers and uses the concept of 'factorized convolution' which splits the convolutional layer into two separate layers - a 1X1 convolution layer and a large convolutional layer.

Selected Approach

VGG16 requires the input image size to be (224, 224). The transfer learning technique is used in this algorithm. The base model used is VGG16 followed by a dense layer and a dropout layer. Relu and Softmax are the activation functions used in this model. After trying various combinations of hyperparameters the best combination is as follows:

Sr.No.	Parameter	Value
1.	Batch size	16
2.	No of epochs	30
3.	Learning rate	0.0001
4.	Optimizer	Adam
5.	Performance matrix	Accuracy

Table 6.3: VGG16: Hyperparameters

Chapter 7

SOFTWARE TESTING

7.1 Type of Testing

7.1.1 Alpha Testing

To check if the product meets the business requirements and functions accurately, Alpha testing can be carried out. It is the first end-to-end testing strategy, specifically performed by employees in a working environment. This method is typically used to check if the product works the way it is expected to.

Alpha testing has been carried out on the system by the team members, who are working and contributing in the same project. Testing has been done by taking various possible test cases into consideration. A few of those test cases have been mentioned below. The system could successfully pass all the test cases and hence reaches the accuracy of almost 99%. No bugs have been encountered throughout the testing process and are not expected to occur in near future. Moreover, both the facilities provided on the website (Self-diagnostic Test and Ultrasound Image Test) produces accurate results. The results are conveyed to the tester in a clearly readable and understandable manner. Since, we being the developers of the web site and have knowledge about the internal design, it can be stated that 'white-box' testing is also successfully completed on the afore mentioned system.

In conclusion, till now no issues and challenges have been faced by us as developers.

7.1.2 Beta Testing

To check the functionality of a system in real life, Beta Testing is performed. This type of testing is done by the potential users who are actually going to be the consumers of the website.

Test results are analyzed on the basis of the feedback provided by these users.
Beta testing has been carried out on our system. We had selected our family members, friends and family doctors as the users and were asked to make use of the website. After using the website, they were asked to provide the feedback on the functionality of the features provided. Additionally, they were asked to suggest advanced facilities from the users' point of view. After considering the feedback, it has been noted that the current working environment with provided functionality is working properly. Users are able to access the website easily and have encountered no difficulty in its use. Advanced facilities like Chat bots have been suggested by the users, which we consider as the future scope for the project.
In short, the system is working fairly good from users' perspective.

7.2 Test Cases and Results

Test Case ID	Test Objective	Expected Result	Actual Result	Test Status
TC01	Launch the web page.	Home page must be displayed.	Successful display of home page.	Pass
TC02	Working of Navigation Bar	All options must redirect to the particular page on click	Navigation Bar redirecting the user to a particular page on click	Pass
TC03	User Manual Video	Video must start after clicking on User Manual Link	Video starts after clicking on User Manual Link	Pass
TC04	User Manual Video uploaded on the third-party website	User must not face the issues when redirect to the third-party website	Video starts successfully without any disturbance	Pass
TC05	User entering incomplete data	Form must show error about the incomplete form status	Browser Popup (Alert Popup) will be displayed	Pass

Test Case ID	Test Objective	Expected Result	Actual Result	Test Status
TC06	Checking the type of image uploaded by the user	Test result if the format is correct	Display of test result page	Pass
TC07	Checking the type of image uploaded by the user	Error if the image is not of the required type	Error due to wrong image type	Pass
TC08	Displaying the test result	Test result must be displayed on the screen as 'You have PCOS' or 'You don't have PCOS'	Display of accurate on the screen	Pass
TC09	Nearest Gynecologists suggested by system	Suggestion of the nearest possible Gynecologist	List of Gynecologists in the vicinity	Pass
TC10	Healthy Life Style page must load	If the result is 'No PCOS' user can click on the link to navigate to healthy lifestyle page	Successful loading of the healthy lifestyle page	Pass

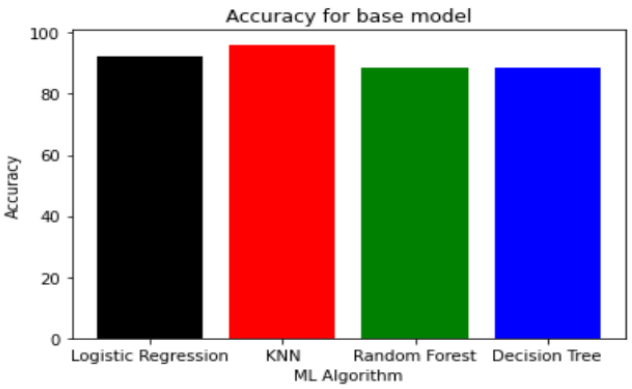
Chapter 8

Results, Analysis, and Screenshots

8.1 Performance comparison

8.1.1 Semi-supervised learning

The accuracy of KNN (K nearest neighbor) was found to be the highest at 96.15%, followed by Logistic Regression at 92.31%, Random Forest Classifier at 88.46%, and Decision Tree Classifier at 88.46%. Therefore, the KNN algorithm was selected to build the final model for predicting labels for an unlabeled dataset, as it provided the best results with the highest accuracy.



	Algorithm	Accuracy
1	KNN	96.15
0	Logistic regression	92.31
2	Random Forest Classifier	88.46
3	Decision Tree Classifier	88.46

Figure 8.1: Base model comparison

Upon combining the data and retraining, the KNN algorithm showed the highest accuracy that is 97.08%, the accuracy of Logistic Regression is 94.12%, the accuracy of the Random forest classifier is 88.24%, Decision tree classifier showed 88.24% accuracy.

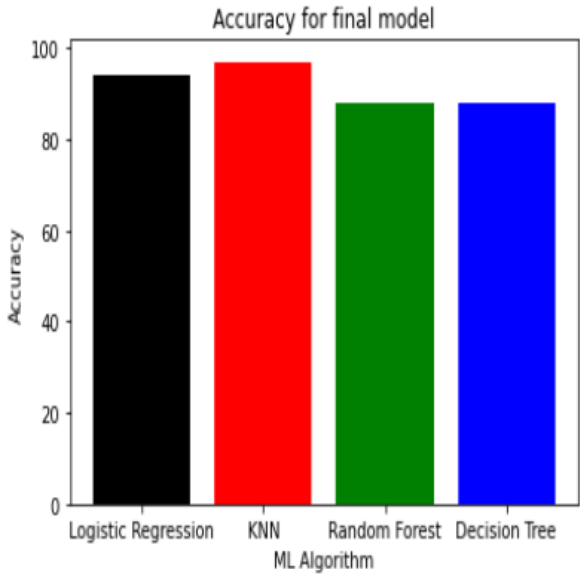


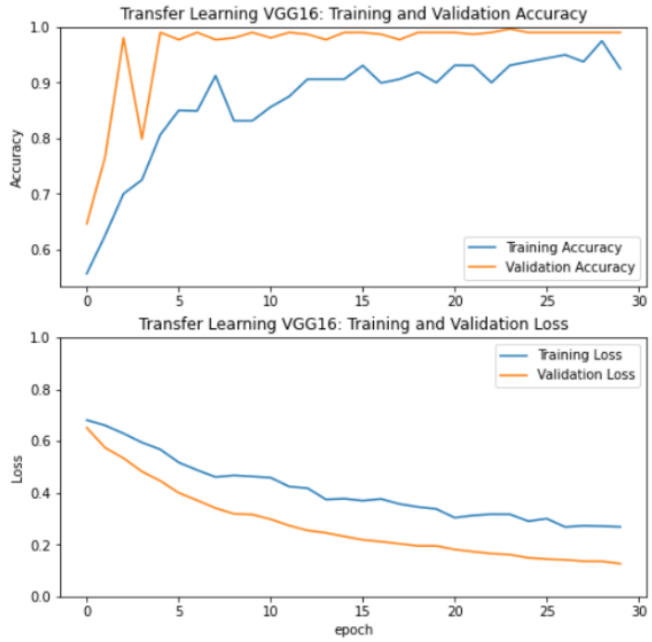
Figure 8.2: Accuracy comparison for final model

	Algorithm	Accuracy
1	KNN	97.06
0	Logistic regression	94.12
2	Random Forest Classifier	88.24
3	Decision Tree Classifier	88.24

Figure 8.3: Retrained models and their accuracies

8.1.2 Deep Learning

After thorough experimentation, the VGG16 achieved the highest accuracy of 98.96% followed by Inception V3 with an accuracy of 97.1%, and at third position, ResNet 50 with an accuracy of 95%.




```
25/25 [==============================] - 2s 65ms/step - loss: 0.1321 - accuracy: 0.9896
Model accuracy on test set: 99.0%
```

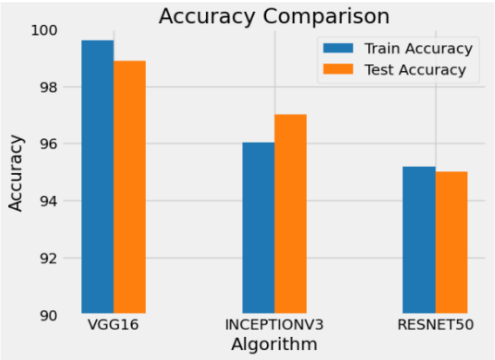
Figure 8.4: Accuracy and loss graph - VGG16


```
25/25 [==============================] - 3s 57ms/step - loss: 0.2647 - accuracy: 0.9506
Model accuracy on test set: 95.1%
```

Figure 8.5: Accuracy and loss graph - ResNet50


```
25/25 [==============================] - 3s 68ms/step - loss: 0.2816 - accuracy: 0.9714
Model accuracy on test set: 97.1%
```

Figure 8.6: Accuray and loss graph - InceptionV3




```
accuracy_df
```

	Algotihm	Train Accuracy	Test Accuracy
0	VGG16	99.6	98.9
1	INCEPTIONV3	96.03	97.01
2	RESNET50	95.19	95

Figure 8.7: Comparison of VGG16, ResNet50 and InceptionV3

8.2 Result

For the ML module, we selected KNN as the base as well as the retraining model. The model obtained an accuracy of 97.08% on test data. For DL module, we selected VGG16 architecture with an accuracy of 98.96% for our system.

KNN:	precision	recall	f1-score	support
0	0.97	1.00	0.98	31
1	1.00	0.67	0.80	3
accuracy			0.97	34
macro avg	0.98	0.83	0.89	34
weighted avg	0.97	0.97	0.97	34

Figure 8.8: KNN - Classification report

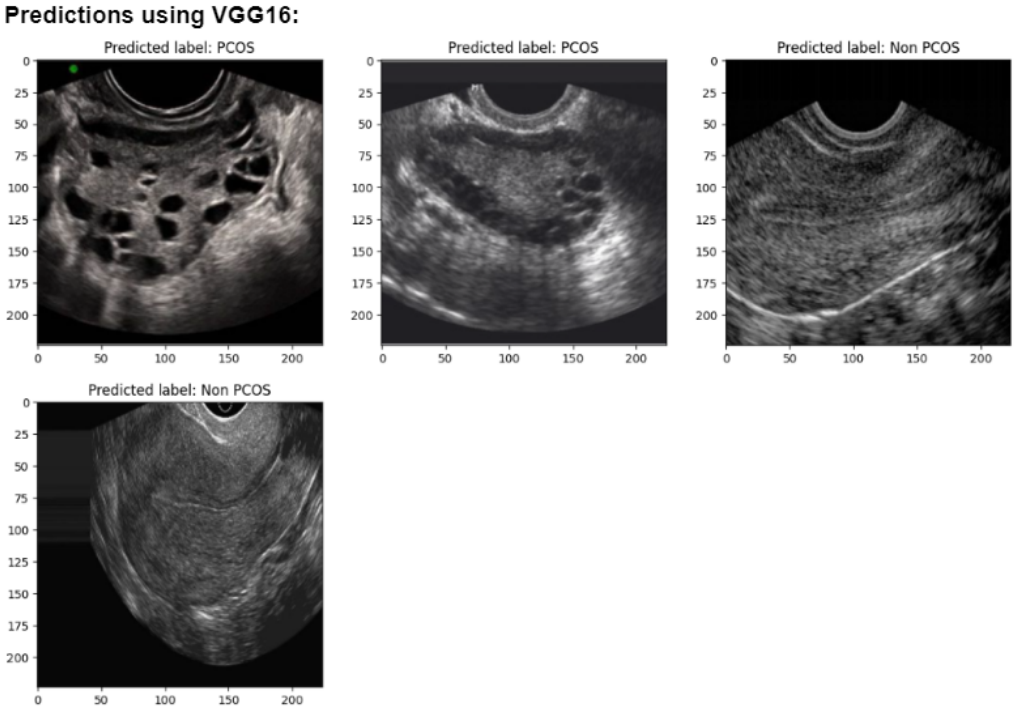


Figure 8.9: Predictions using VGG16

8.3 Screenshots

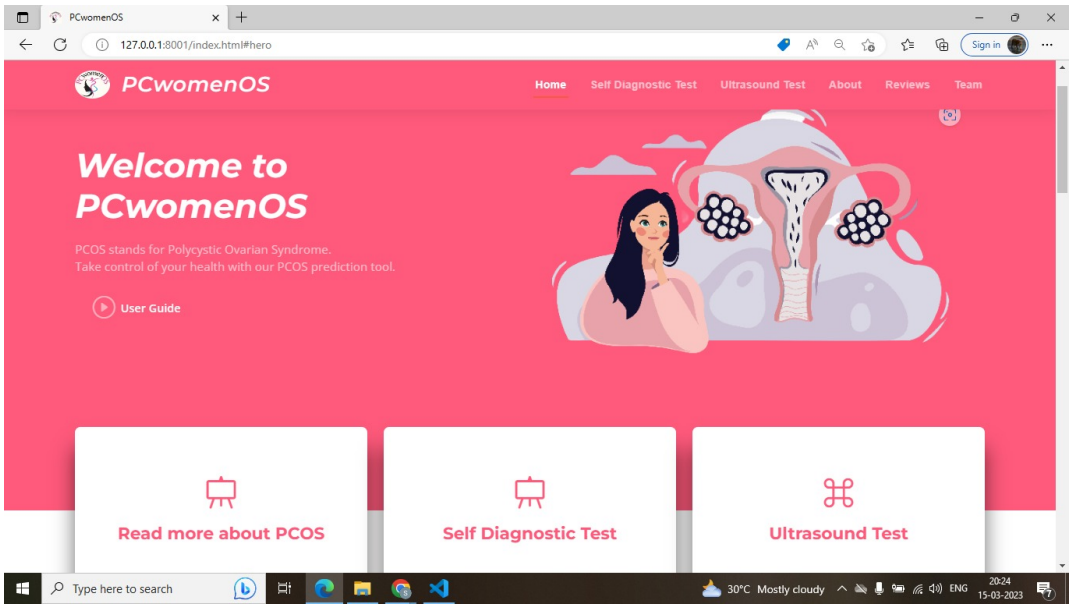


Figure 8.10: Home Page

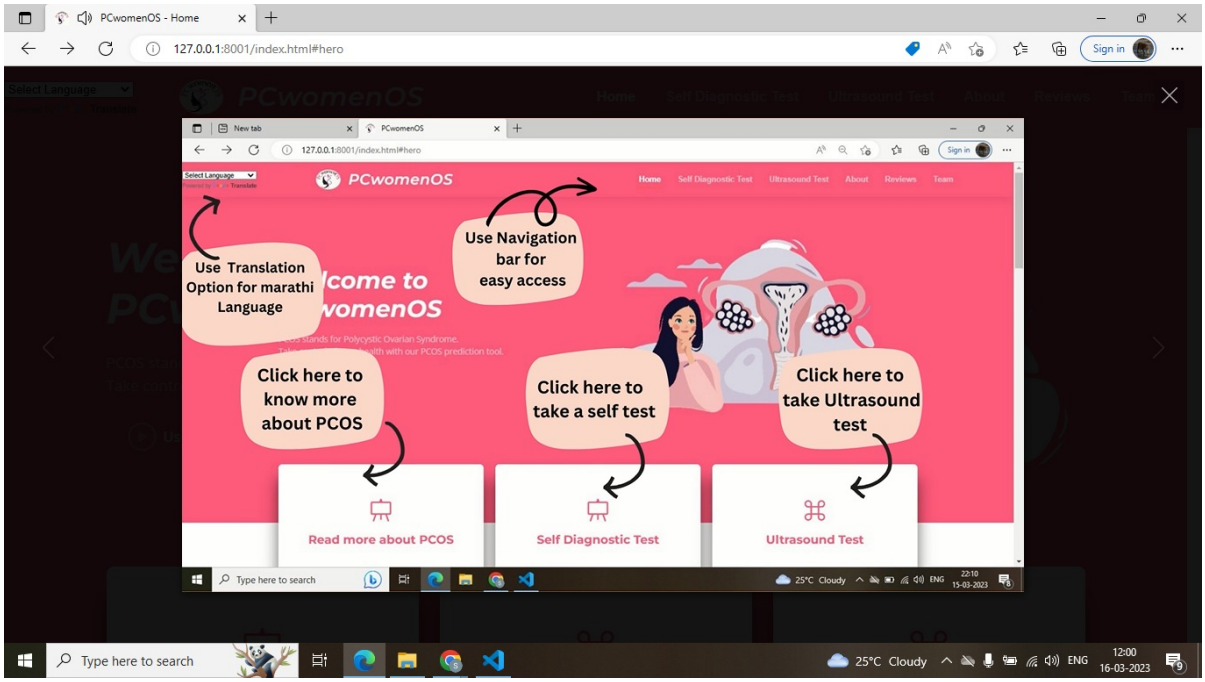


Figure 8.11: User Manual

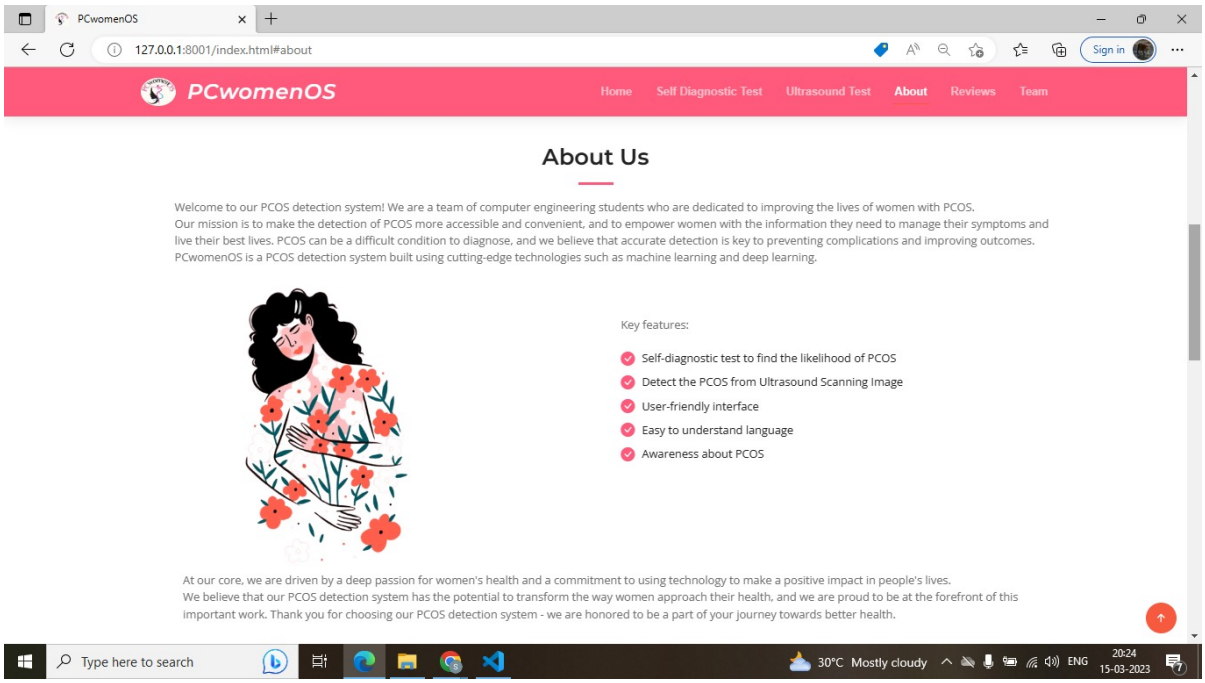


Figure 8.12: About Us

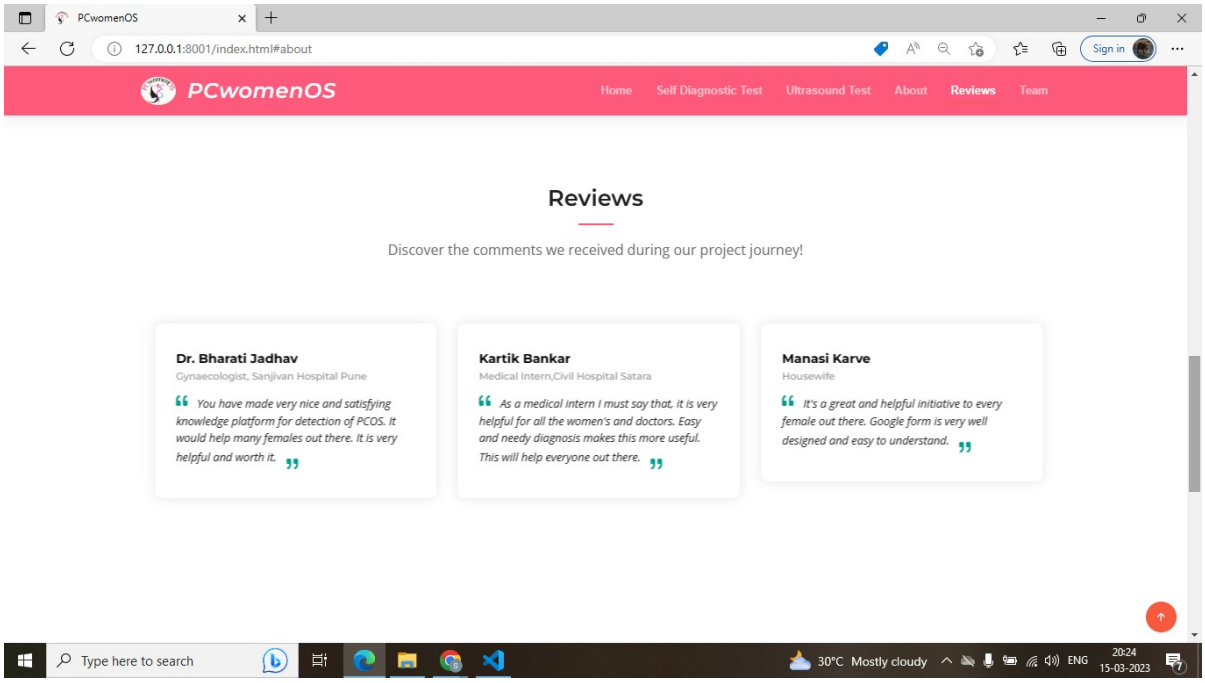


Figure 8.13: Reviews

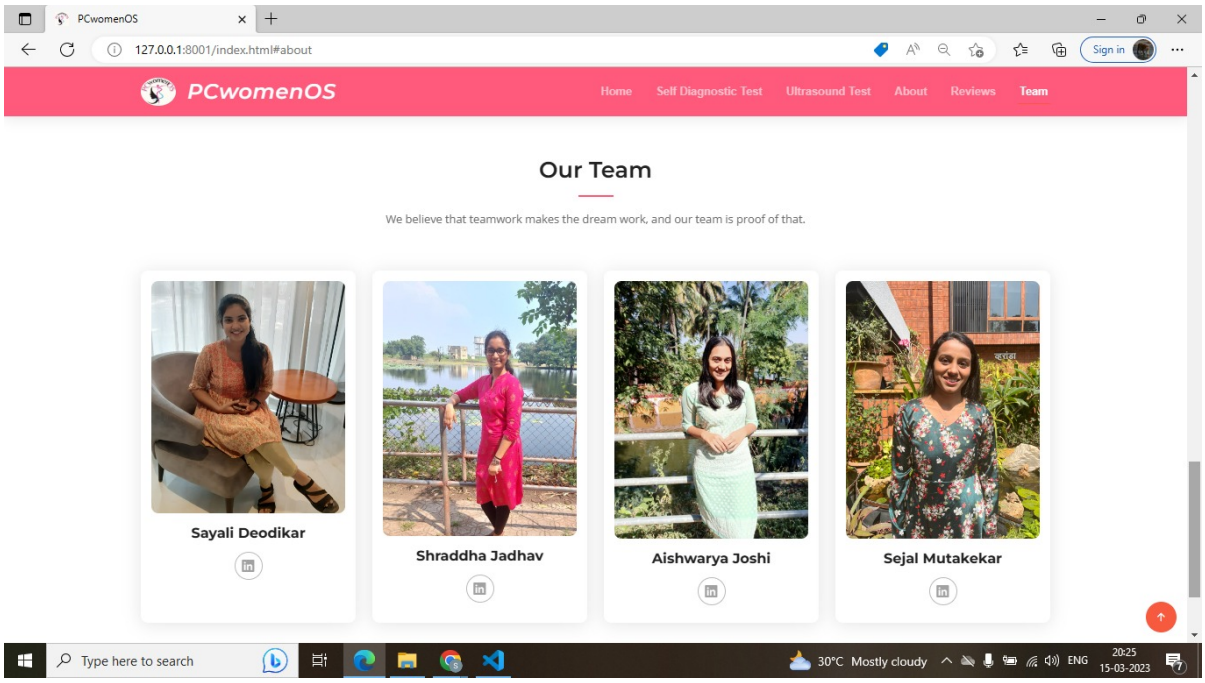


Figure 8.14: Our Team

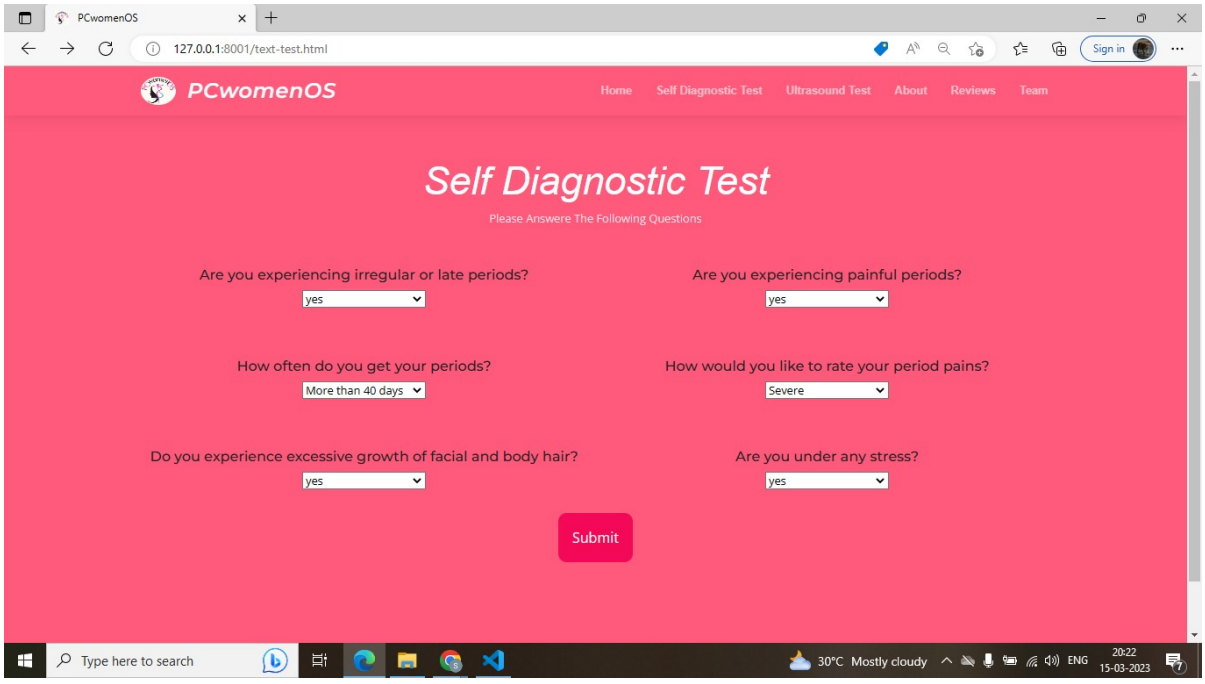


Figure 8.15: Diagnostic Test using the text data entered by the User (Part – I)

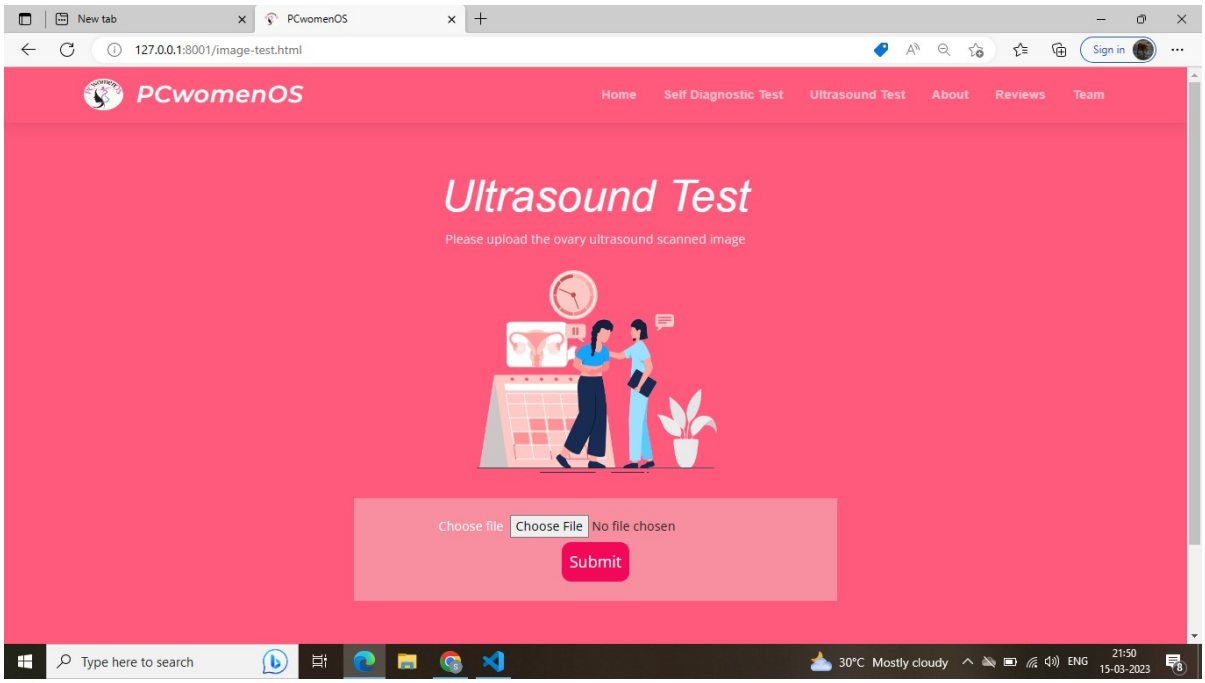


Figure 8.16: Diagnostic Test using the Ultrasound Image uploaded by the User

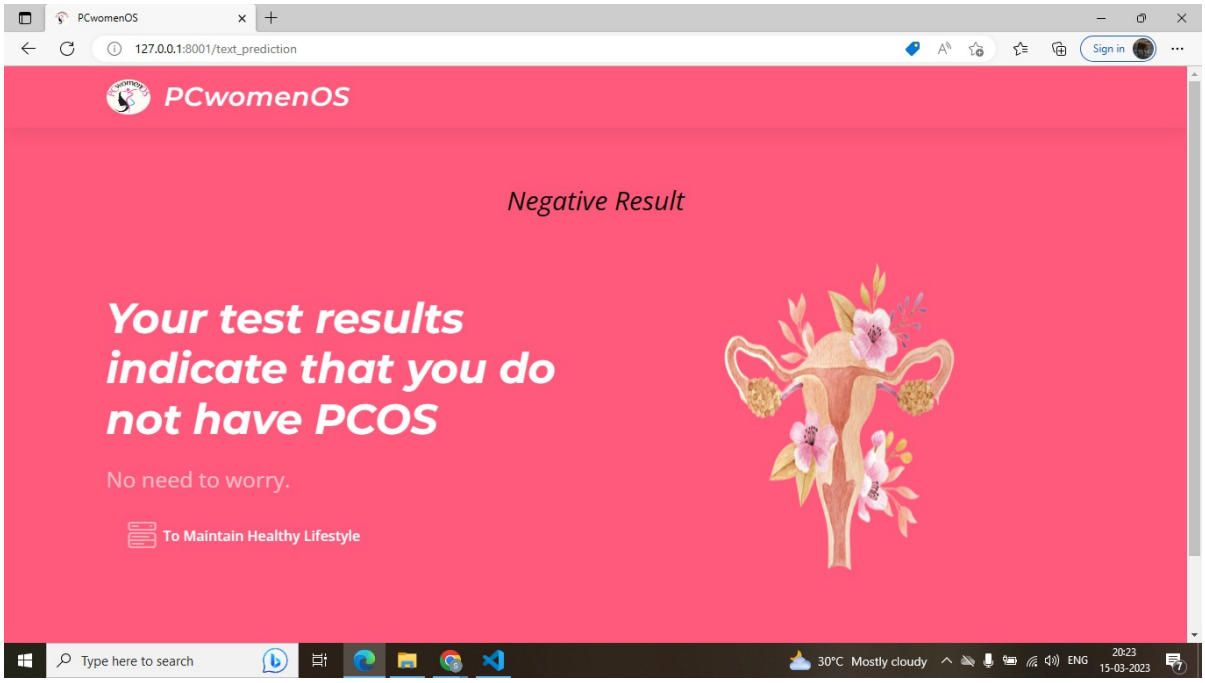


Figure 8.17: Diagnostic Test Result as 'No PCOS detected'

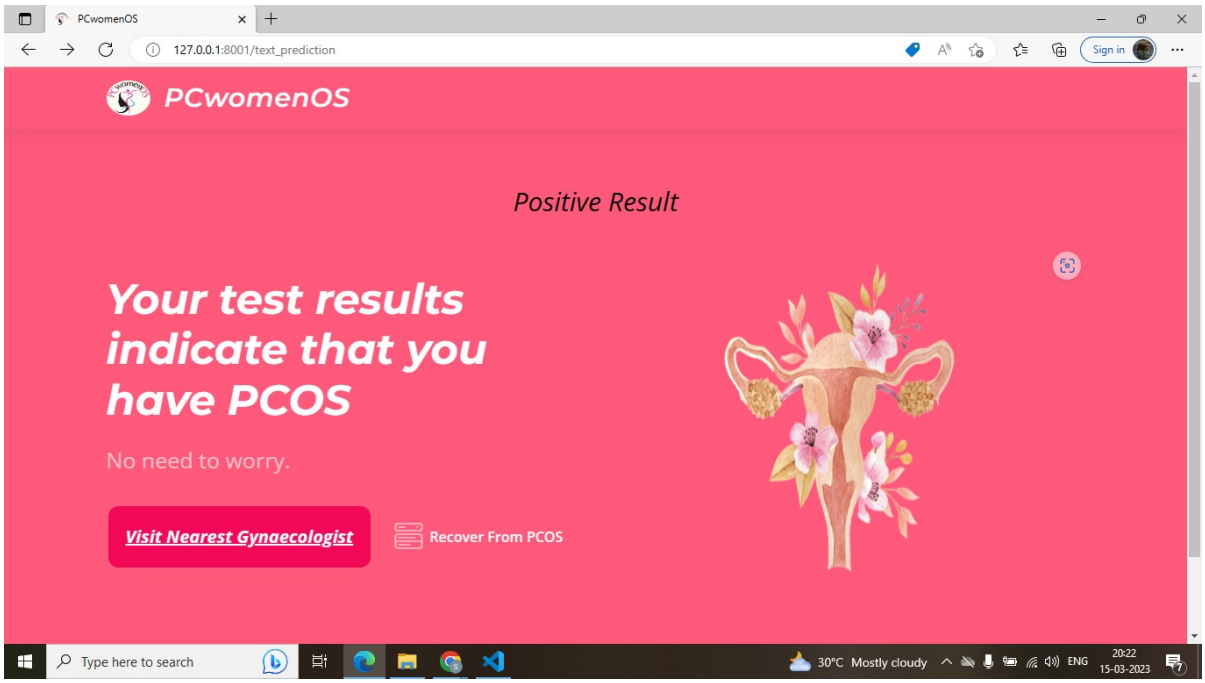


Figure 8.18: Diagnostic Test Result as 'PCOS detected'

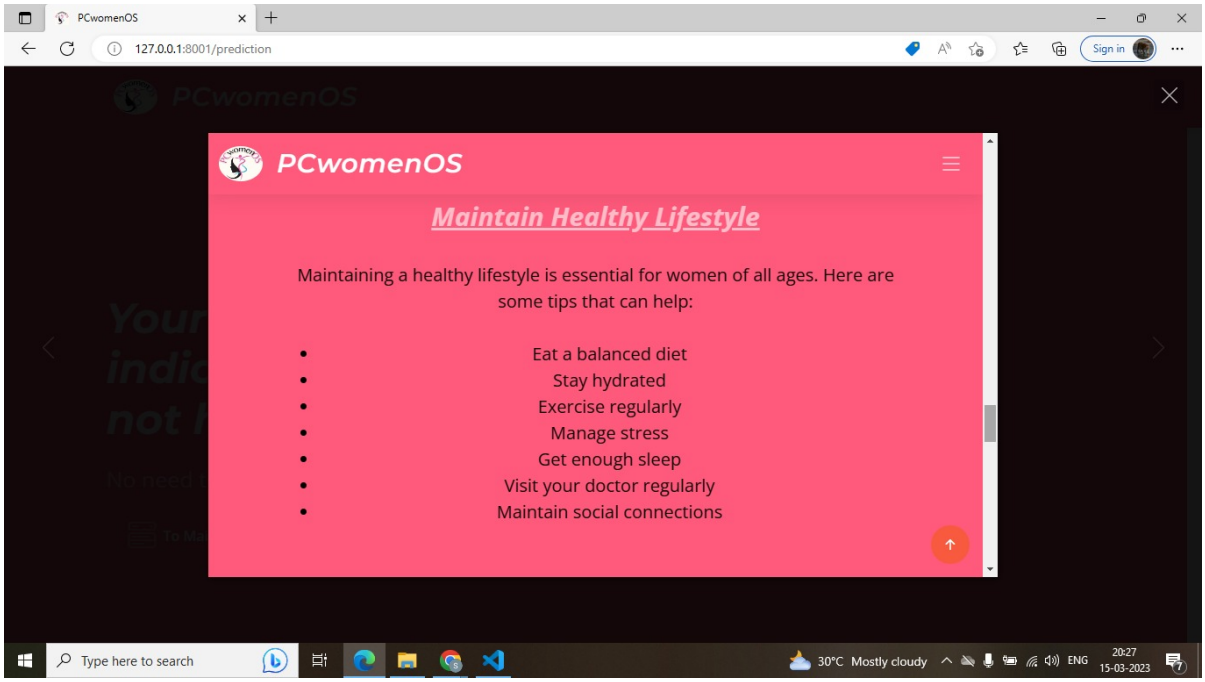


Figure 8.19: Maintain Healthy Lifestyle

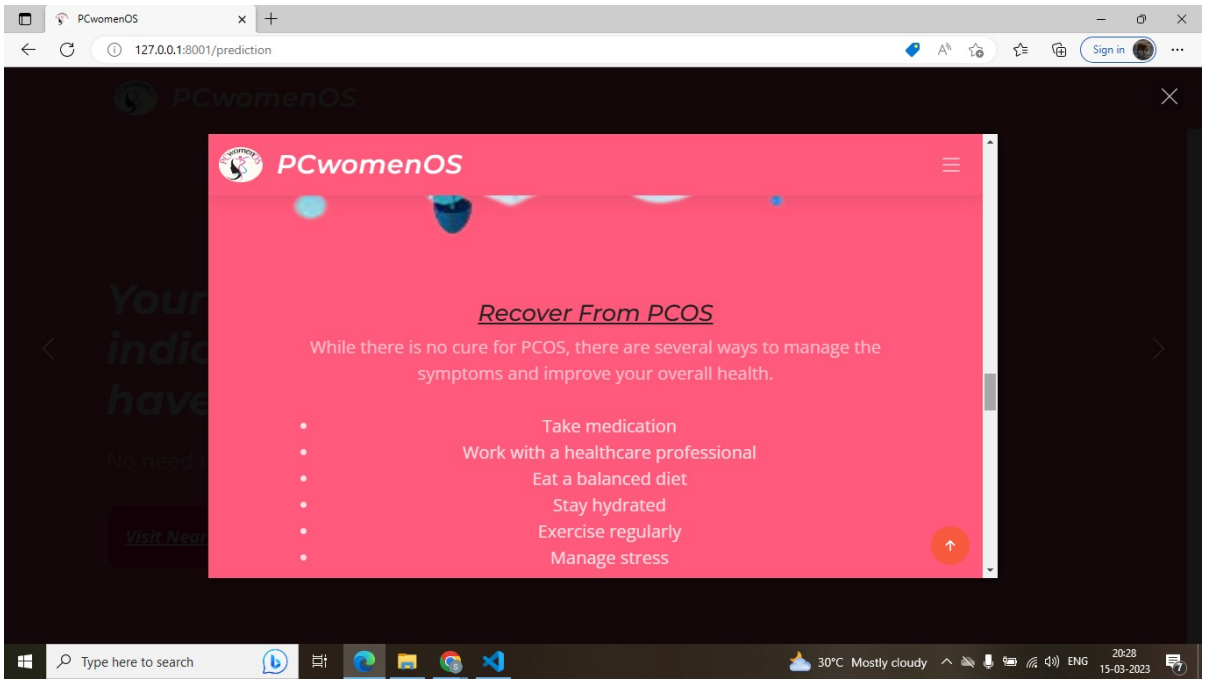


Figure 8.20: Treat PCOS

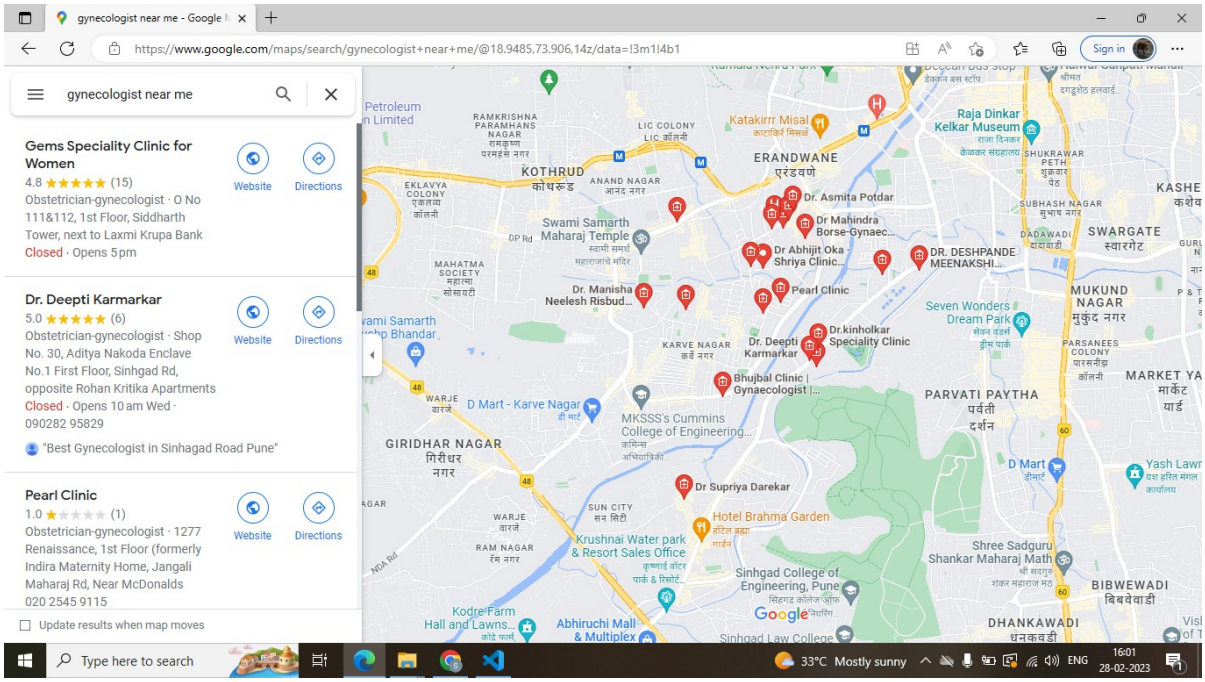


Figure 8.21: Nearest Gynaecologists Recommendation

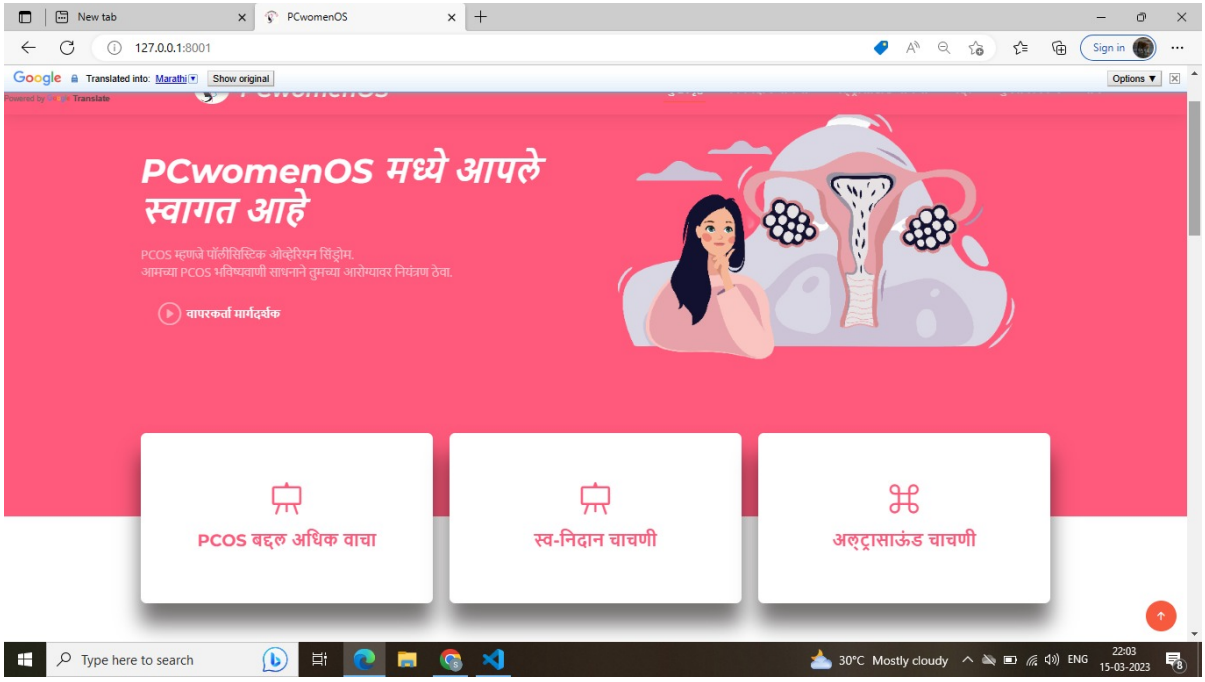


Figure 8.22: Multiple Language Translator For Webpages

Chapter 9

CONCLUSIONS

9.1 Conclusion

During these project data science life cycle was followed starting from problem identification ,literature review was studies ,design of system architecture was done ,chose the accurate ML model for text data processing and CNN model for image data processing, suggested the nearest gynecologist if the user got detected with pcos and some other helps are provided such as how to maintain the healthy lifestyle , what are the natural remedies to recover from the pcos are suggested to user. Multiple machine learning algorithms are used to train the model for text modules such as logistic regression, random forest,decision tree, naive bayes and KNN algorithm . Out off all these algorithms KNN gives the highest accuracy of 97.06% and it was used for prediction purposes . For the ultrasound image test module VGG- 16,resnet model is used in which VGG16 gives the accuracy of 95% and it was selected for image module prediction. By completing all the above mentioned modules an end goal of the project is achieved.

9.2 Future Work

Making the website for a build system restricts its scope. Nowadays. Not only the website but also the android applications have gained a huge popularity and it is also easily handleable by the user. Android applications are now used by every generation (younger to older). More and more people are now preferring cell phones so that people may prefer to use an android application over a website. Therefore, developing an android application makes the system easily available and for getting quick and accurate diagnosis within a short time.

As an additional contribution to this creating a chatbot which could answer the queries of user related to PCOS , if the queries are not resolved by the chatbot assistance then connecting the call to doctors so that user can get satisfactory results to their query within low budget and it saves the time for going to the hospital .By storing the data inputs of the user and their prediction we can use this additional data information for training the model which may increase the performance and accuracy of each module.

9.3 Applications

1. Quick and easy diagnosis of PCOS : Its not possible for an individual every time to visit the hospitals may be due to the busy schedules or one may not even afford paying for the disgnosis. In this case, this platform will help in easy diagnosis at almost zero cost. Moreover, the results will be instant.

2. Recommendation of nearest doctors to treat PCOS : The system will recommend the user the nearby places, so that the patient can quickly treat the disorder.

3. Suggestions to maintain healthy life style : Its a good things, if a person is not suffering from PCOS. But, its a resposibility of every individual to maintain the healthy lifestyle to continue living the similar healthy life further.

4. Health Camps : This system can be used in the Health Camps, where one can spread awarness among the people and help them to disgnose and treat this disorder.

APPENDIX A

Feasibility check

Operational Feasibility

The system is user-friendly and simple to use. A user does not need to have advanced knowledge to deal with the system.

Technical Feasibility

1. With good Internet connectivity, the system will perform efficiently and provide fast results.
2. No issues regarding legality will be faced as the libraries and software used are open-source and free to use.

Problem Analysis using Algebra

Algorithmic Techniques:

- Trend Identification: Number of patterns available in the dataset n It is time-consuming to retrieve the information if n is greater than I. Hence, the time complexity of this algorithm is O(n). This mathematical model is NP-Complete.
- Trend monitoring: The project is NP-Complete as the output can easily be determined and it will reach the final step in the month of April with complete implementation and desired output ready, so it is NP-Complete.

Problem Type

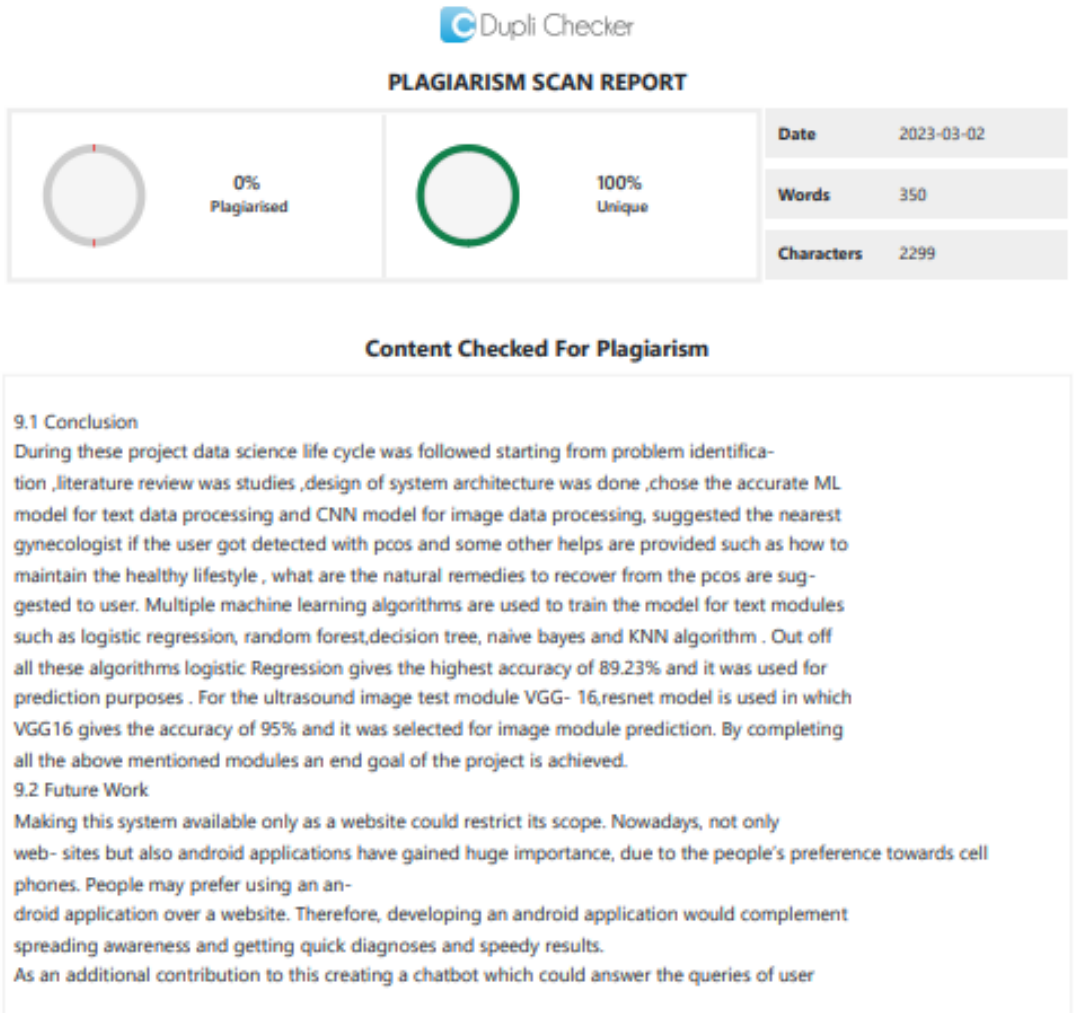
This problem gives a solution in polynomial time. Hence, this problem statement is a P-type problem.

APPENDIX B

Paper Publication Details

- Title: PCOS Detection: A Study of the Literature
- Journal : IJRASET (Internal Journal for Research in Applied Science and Engineering Technology)
- Year : 2023
- Status : Published

APPENDIX C



Dupli Checker

PLAGIARISM SCAN REPORT

0% Plagiarised

100% Unique

Date	2023-03-02
Words	350
Characters	2299

Content Checked For Plagiarism

9.1 Conclusion
During these project data science life cycle was followed starting from problem identification ,literature review was studies ,design of system architecture was done ,chose the accurate ML model for text data processing and CNN model for image data processing, suggested the nearest gynecologist if the user got detected with pcos and some other helps are provided such as how to maintain the healthy lifestyle , what are the natural remedies to recover from the pcos are suggested to user. Multiple machine learning algorithms are used to train the model for text modules such as logistic regression, random forest,decision tree, naive bayes and KNN algorithm . Out off all these algorithms logistic Regression gives the highest accuracy of 89.23% and it was used for prediction purposes . For the ultrasound image test module VGG- 16,resnet model is used in which VGG16 gives the accuracy of 95% and it was selected for image module prediction. By completing all the above mentioned modules an end goal of the project is achieved.
9.2 Future Work
Making this system available only as a website could restrict its scope. Nowadays, not only web- sites but also android applications have gained huge importance, due to the people's preference towards cell phones. People may prefer using an android application over a website. Therefore, developing an android application would complement spreading awareness and getting quick diagnoses and speedy results.
As an additional contribution to this creating a chatbot which could answer the queries of user

Figure 9.1: Plagiarism Check

Bibliography

[1] Sayma Alam Suha Muhammad Nazrul Islam. "An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image", Scientific Reports, 2022

[2] Irteza Enan Kabir A.K.M. Salman Hosain, Md Humaion Kabir Mehedi. "PCOnet: A convolutional neural network architecture to detect polycystic ovary syndrome (PCOS) from ovarian ultrasound images" 2022.

[3] Kashif Munir Ali Raza Shazia Nasim, Mubarak Almutairi, and Faizan Younas. "A novel approach for polycystic ovary syndrome prediction using machine learning in bioinformatics", 10, 2022.

[4] Rekha Radhakrishnan Sumalatha P. Subha R, Nayana B R. "Computerized diagnosis of polycystic ovary syndrome using machine learning and swarm intelligence technique", 2022.

[5] Hyunsun Lee. Angela Zigarelli, Ziyang Jia. "Machine-aided self-diagnostic prediction models for polycystic ovary syndrome: Observational study" 2022.

[6] Prof. Dr. Mrs. Suhasini A. Itkar Kinjal Raut, Chaitrali Katkar. "PCOS detect using machine learning algorithm", 09, 2022.

[7] Arun Shivsharan. Shubham Bhosale, Lalit Joshi. "Pcos (polycystic ovarian syndrome) detection using deep learning" 04, 2022.

[8] Rongxin Fu Xue Lin Ya Su Xiangyu Jin Han Yang Xiaohui Shan Wenli Du Qin Huang Hao Zhong Kai Jiang Zhi Zhang Lina Wang Wenqi Lv, Ying Song and Guoliang Huang. "A deep learning algorithm for automated detection of polycystic ovary syndrome using scleral images", 2022.

[9] Siji Jose Pulluparambil 1 Subrahmanya Bhat. "Medical image processing: Detection and prediction of PCOS – a systematic literature review". 5, 2022. ISSN 2581-6411

[10] Ashwini Kodipalli and Susheela Devi. “Prediction of PCOS and mental health using fuzzy inference and SVM” 9, 2021.

[11] Durgesh Nandan Anurag Mahajan. Shruti Bhargava Choubey, Abhishek Choubey. “Polycystic ovarian syndrome detection by using two-stage image denoising”, 38, 2021.

[12] Vineesha K. Vikas B, Radhika Y. Detection of polycystic ovarian syndrome using convolutional neural networks. 13, 2021.

[13] Sweta Kumari. “Classification of PCOS/PCOD using transfer learning and gan architectures to generate pseudo ultrasound images” 2020.

[14] Namrata Tanwani. “Detecting PCOS using machine learning”, 07, 2020. ISSN 2348-3121.

[15] Holger H. Hoos Jesper E. van Engelen1. “A survey on semi-supervised learning”,2019