6COSC023W

Computer Science Final Project

Project Proposal and Requirements Specification (PPRS)

Handout

Web Application for Diagnosing PCOS Patients Using Machine Learning

Student: Rachel Naadia Cooray (19564441)

Supervisor: Dr Malarvizhi Kaniappan Chinnathai

Degree: BSc (Hons) Computer Science

School of Computer Science & Engineering

College of Design, Creative and Digital Industries

University of Westminster

14th November 2024

Declaration

This document has been prepared based on my own work. Where other published and unpublished source materials have been used, these have been acknowledged in references.

Student Name: Rachel Naadia Cooray

Date of Submission: 14th November 2024

Table of Contents

List of Figures

List of Tables

1. Aims, Objectives and Scope

1.1 Research Question

How accurately can a Machine Learning (ML) based web application predict the likelihood of Polycystic Ovary Syndrome (PCOS) in women using patient-provided medical and hormonal data?

1.1.1 Aim

The primary aim of this project is to develop a web-based application that offers preliminary assessments for the likelihood of PCOS using machine learning. By using patient-provided medical and hormonal data, this project aims to create an accessible, user-friendly platform to support individuals and healthcare professionals seeking an initial risk evaluation for PCOS. The application will provide a combination of prediction capability and educational resources to inform users and support awareness around PCOS symptoms, causes, and management.

1.2 Objectives

- To develop a user-friendly interface for users to input their medical and hormonal data, with guided descriptions to improve usability.
- To integrate key features such as:
    1. A simple data-entry form to collect relevant health information.
    2. A machine learning-based classification model to assess PCOS risk.
    3. Data visualization tools (e.g., charts and graphs) to help users interpret their prediction results.
    4. An informational page with credible resources on PCOS symptoms, causes, and management for educational purposes.
- To ensure adaptability of the application for various devices, delivering a seamless user experience across platforms.
- To provide clear, accessible feedback on PCOS risk predictions, supplemented by visual elements to enhance user comprehension.

1.3 Target Audience

- **Primary audience**
  Individuals who believe they may be at risk of PCOS and are seeking a preliminary risk assessment.

- **Secondary audience**
  Healthcare providers interested in using a machine learning-based tool for initial PCOS risk detection and patient education.

1.4 Scope

1.4.1 Inclusions

This project will include the following key features:

- A **data input form** for secure and anonymous collection of user-provided medical and hormonal data, which will be used solely for predictive purposes.
- A **machine learning model** to analyze input data and predict the likelihood of PCOS based on trained classification models.
- **Results visualization tool** that display prediction outcomes and other relevant statistics, allowing users to understand the basis of the prediction.
- An **information page on PCOS** with credible resources that cover PCOS symptoms, causes, and management tips, fostering user education and awareness.

1.4.2 Exclusions

This project will exclude the following:

- Any storage of personally identifiable or sensitive health data. User data is anonymous and not stored beyond the prediction session.
- Clinical diagnosis capabilities, as the tool provides only a preliminary assessment and should not be used as a substitute for professional medical evaluation.
- Integration with healthcare systems or electronic medical records (EMRs), as it is intended as a standalone web-based tool.
- User login or long-term data storage for user tracking or personalization, since the application does not retain data post-session.

1.4.3 Limitations

- The prediction accuracy is limited by the quality and scope of the training data, which may affect the generalizability of results for diverse user populations.
- The application is intended solely for preliminary assessments and is not a substitute for clinical diagnosis or professional medical advice.

2. Definition of the Problem

2.1 Problem Description

Among women of reproductive age, PCOS is a common hormonal disorder and around 70% of them remain undiagnosed (Adla et al., 2021). Ovarian cysts, irregular menstruation cycles, and high testosterone levels are its defining characteristics. As it often results in symptoms including acne, excessive hair growth, obesity, and infertility, this disorder has a significant impact on women's health (Hdaib et al., 2022).

Although being common, PCOS can be challenging to diagnose, requiring a number of tests and professional consultations. Many women, especially in areas with poor healthcare infrastructure, are either unaware of the disorder or may not receive treatment because they have limited access to healthcare services.

Although lifestyle and genetic factors are thought to play a part, the precise origin of PCOS is yet unknown (Gandhi, Patel and Dave, 2024).

2.2 Significance

Addressing this issue is important for healthcare technology as it helps with early detection of PCOS, which could lead to better health outcomes for those who are impacted. In order to enable users to seek expert medical assistance sooner, an easily available online PCOS prediction tool can serve as a valuable resource for initial assessment.

The project further contributes to the field by evaluating how well machine learning algorithms predict PCOS risk, which may lead to the development of related applications for other under-diagnosed medical disorders. The overall goal of this is to improve health accessibility through data-driven and digital solutions.

2.3 Proposed Solution

This project aims to close the diagnosis gap by developing a web-based tool that applies machine learning to provide a preliminary PCOS risk assessment based on user entered hormonal and medical data. This user-friendly tool aims to help larger health accessibility initiatives by encouraging consultation with specialists and promoting early detection.

2.4 Context Diagram

The diagram below shows how the system and its users interact with each other.

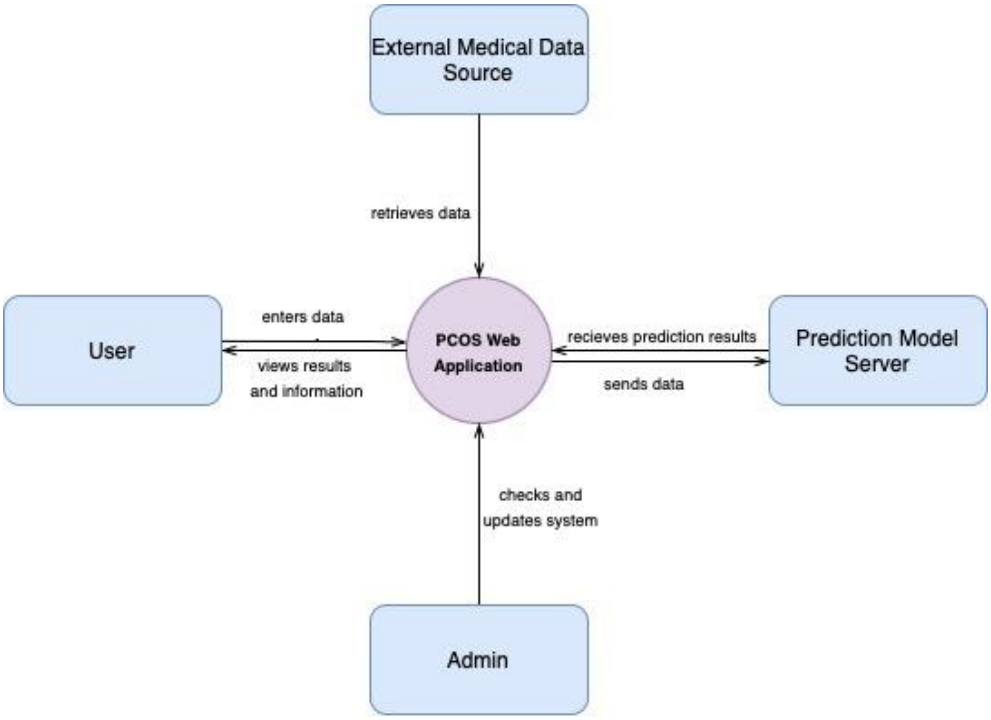


Figure 1: Context Diagram (Self composed)

3. Background Review

3.1 Key Findings and Research Gaps

The existing literature on PCOS prediction studies offers important insights into how machine learning algorithms are applied and how well they perform.

Using the **filter approach**, Gandhi, Patel, and Dave (2024) chose 10 significant features from a total of 44, choosing features that had the highest ranks according to univariate metrics.

Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Support Vector Machine (SVM) were among the **classification techniques** they reviewed. The Random Forest Classifier achieved the best accuracy of 83.48% for PCOS identification, according to their findings.

The **variations in datasets and approaches** are shown by the notable variance in algorithm performance amongst research. In their study, Tiwari et al. (2022) also used Random Forest and achieved an accuracy of 93.25%.

The study also reveals a weakness in the **lack of using clustering** methodologies, which may be able to reveal PCOS subtypes for more in-depth understanding (Gandhi, Patel, and Dave, 2024).

While Decision Tree algorithms show excellent accuracy in identifying PCOS risk, more effective predictive models that **combine medical and lifestyle characteristics** are still needed, according to the findings of Priyadharshini et al. (2024).

In addition to **expert consultation** with an OB-GYN, Adla et al. (2021) used the **Analysis of Variance** (ANOVA) test for feature selection.
They looked at a broader range of classification methods, such as AdaBoost, K-Nearest Neighbours, and Naïve Bayes. Using 10-fold cross-validation, the LSVM (Linear Support Vector Machine) produced an accuracy of 91.6%, precision of 93.66%, and recall of 80.6%.

The **weaker recall** suggests the tendency to fail to recognise some actual positives despite its high precision. This implies that the best possible balance between recall, accuracy, and precision must be found.
Furthermore, to improve the model's accuracy, **larger datasets** and additional medical features are required.

High-quality, **standardised datasets** that cover a range of populations are of importance, according to the study by Radhakrishnan et al. (2023).
There is a need for advanced algorithms to differentiate unique features because current feature selection techniques frequently include redundant features that do not improve the diagnosis. Additionally, enhancing **model generalisability** across various demographic groups and gaining clinical acceptability of AI-based diagnostic tools are still difficult tasks.

3.2 Comparison with Similar Software Applications/Products

Application or Product	Strength	Weakness
NHS Website	A trustworthy and easily available source of general PCOS information by offering reliable, medically evaluated information about PCOS symptoms, causes, and treatment options.	Does not offer a predictive tool or personalized assessments for PCOS risk, limiting interactivity and personalized user engagement.
Flo App	Widely used for managing reproductive health because it offers fertility tracking with individualised insights based on the menstrual cycle.	Does not specifically address PCOS prediction or provide detailed information on PCOS management, limiting its application for those seeking focused PCOS insights.
PCOS App	Provides information about PCOS, including its causes, symptoms, and suggested lifestyle. It acts as a useful educational aid.	Lacks a predictive model to assess an individual's risk of developing PCOS.
PCOS Tracker	Tracks menstrual cycles and symptoms for individuals managing PCOS, helping users monitor their health.	Does not provide risk prediction or medical assessments.

Table 1: Comparison with Similar Products

Difference of this project to existing similar products

With the help of medical and hormonal data, this project will provide a unique predictive model to determine a person's risk of PCOS. In contrast to the current products, it integrates PCOS prediction, educational content, and results visualisation. To make it a more complete tool for awareness and early risk assessment, the project will incorporate a data input form, prediction model selection, result visualisation, and an educational information website on PCOS symptoms, causes, and management.

4.Tools (Hardware/Software)

4.1 Hardware

Tool	Purpose
Laptop/PC with 8GB or more RAM	For development, testing, and running machine learning models, to handle data processing and model training tasks efficiently.

Table 2: Hardware Tools

4.2 Software

The table below shows the tools and technologies selected for the implementation of the application, mainly in terms of frontend development, backend development and the machine learning component.

Tool	Resource Type	Purpose
Streamlit	Frontend Framework	Provides an interactive front end for building and deploying the prediction model as a web application with ease.
Flask	Backend Framework	Enables secure backend processing, data handling, and to connect the front end with the prediction model.
Python	Programming Language	Used for developing machine learning models and data analysis.
scikit-learn	Machine Learning Library	Offers essential algorithms for building and evaluating the PCOS classification model.
Pandas and NumPy	Data Manipulation	For data cleaning, pre-processing, and numerical operations needed for efficient model input preparation.
Matplotlib and Seaborn	Visualization Library	To generates visual representations of prediction results, providing clear data insights.

GitHub	Version Control	To manage code versions and repository for project documentation.

Table 3: Software Tools

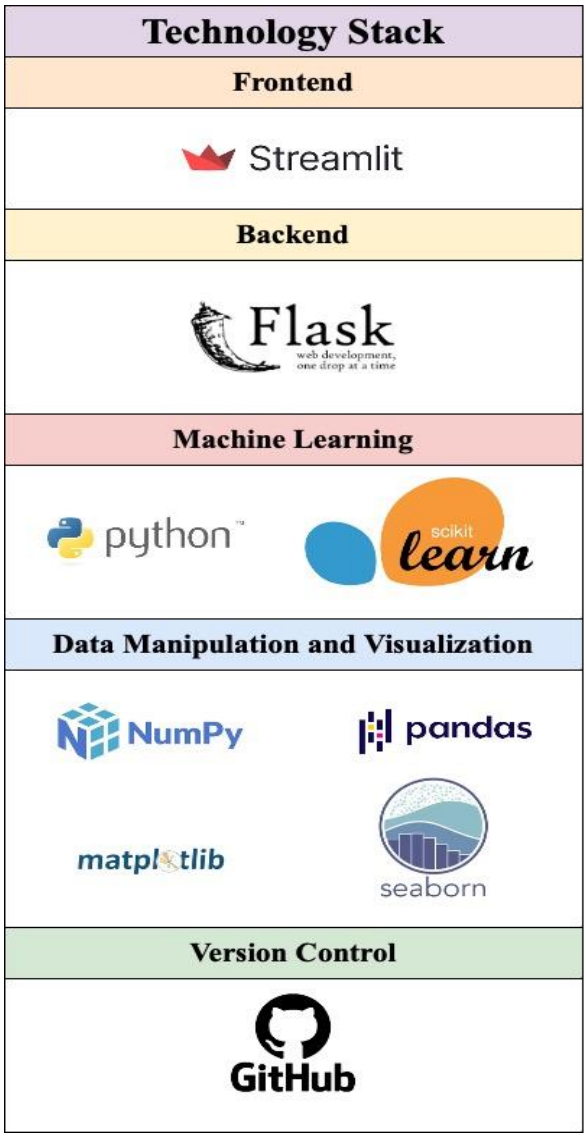


Figure 2: Technology stack (Self composed)

5. Use Cases and Diagrams

5.1 Use Case Descriptions

The tables in this section describe the use cases of this application clearly.

Use Case Name	**Enter Medical Data**	
Use Case ID	**UC-001**	
Description	User enters personal health and medical data required for PCOS prediction.	
Primary Actor	User	
Supporting Actors	System	
Pre-Conditions	User must have access to the website.	
Trigger	User initiates the entry of medical data by selecting the option to start the prediction process.	
Main flow	**Actors**	**System**
	1. User navigates to the data entry section. 3. User enters data and submits it.	2. Displays data input fields for health and medical information. 4. Validates the entered data.
Exception flow	**Actors**	**System**
	2. User corrects errors and resubmits data.	1. If validation fails (e.g., missing fields or invalid data), displays an error message.

Alternate flow	Actors	System
	None	None
Exclusions	None	
Post Conditions	User data is validated and ready for prediction.	

Table 4: Use Case Description 1

Use Case Name	Run Prediction	
Use Case ID	**UC-002**	
Description	System processes the entered data and runs the prediction model.	
Primary Actor	System	
Supporting Actors	User	
Pre-Conditions	User has entered and submitted valid data.	
Trigger	User selects the option to run the prediction.	
Main flow	**Actors**	**System**
	1. User selects “Run Prediction” after entering medical data.	2. Accesses stored prediction model. 3. Processes user data through the model. 4. Generates prediction results.

Exception flow	**Actors**	**System**
	2. User may retry the prediction tool.	1. If the prediction model encounters an error, displays an error message.
Alternate flow	**Actors**	**System**
	None	None
Exclusions	No prediction is run if data validation fails.	
Post Conditions	Prediction results are generated and available for display.	

Table 5: Use Case Description 2

Use Case Name	**View Prediction Results**	
Use Case ID	**UC-003**	
Description	User views the results of the PCOS prediction and any related insights.	
Primary Actor	User	
Supporting Actors	System	
Pre-Conditions	Prediction has been run, and results are generated.	
Trigger	System completes prediction and displays results.	
Main flow	**Actors**	**System**

	2. User reviews the prediction results.	1. Displays the prediction results to the user.
Exception flow	**Actors**	**System**
		1. If results cannot be displayed, provides a message.
Alternate flow	**Actors**	**System**
	None	None
Exclusions	Results are not displayed if the prediction has not been run.	
Post Conditions	User has viewed the prediction results.	

Table 6: Use Case Description 3

Use Case Name	**Read Information on PCOS**
Use Case ID	**UC-004**
Description	User accesses educational information on PCOS provided by the system.
Primary Actor	User
Supporting Actors	System
Pre-Conditions	User must have access to the website.
Trigger	User selects the option to read information on PCOS.

Main flow	Actors	System
	1. User navigates to the PCOS information section.	2. Displays educational content on PCOS, including symptoms, causes, and treatment options.
Exception flow	**Actors**	**System**
		1. If information is unavailable, displays an error message.
Alternate flow	**Actors**	**System**
	None	None
Exclusions	Content is read-only; no user input or interaction affects it.	
Post Conditions	User has accessed educational information on PCOS.	

Table 7: Use Case Description 4

5.2 Diagrams

5.2.1 System Architecture Diagram

The web application will have a 3 tiered architecture.

1. **Presentation Layer**
   This layer includes the Data Input Form, Information Page, and Visualization Page to facilitate user interaction, data entry, and result visualization.

2. **Application Layer**
   The application's core functionality, including the Prediction Model, processes data and delivers PCOS prediction.

3. **Database Layer**
   This layer stores and manages the Dataset, supporting data retrieval and prediction accuracy.

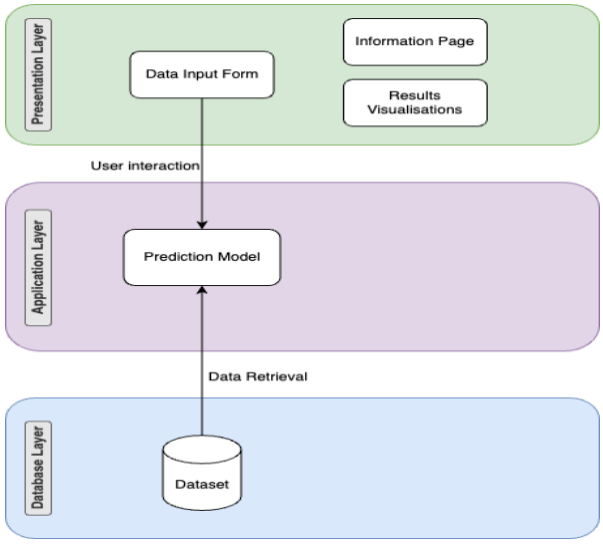


Figure 3: System Architecture Diagram (Self composed)

5.2.2 Use Case Diagram

This diagram shows how the actors interact with the use cases.

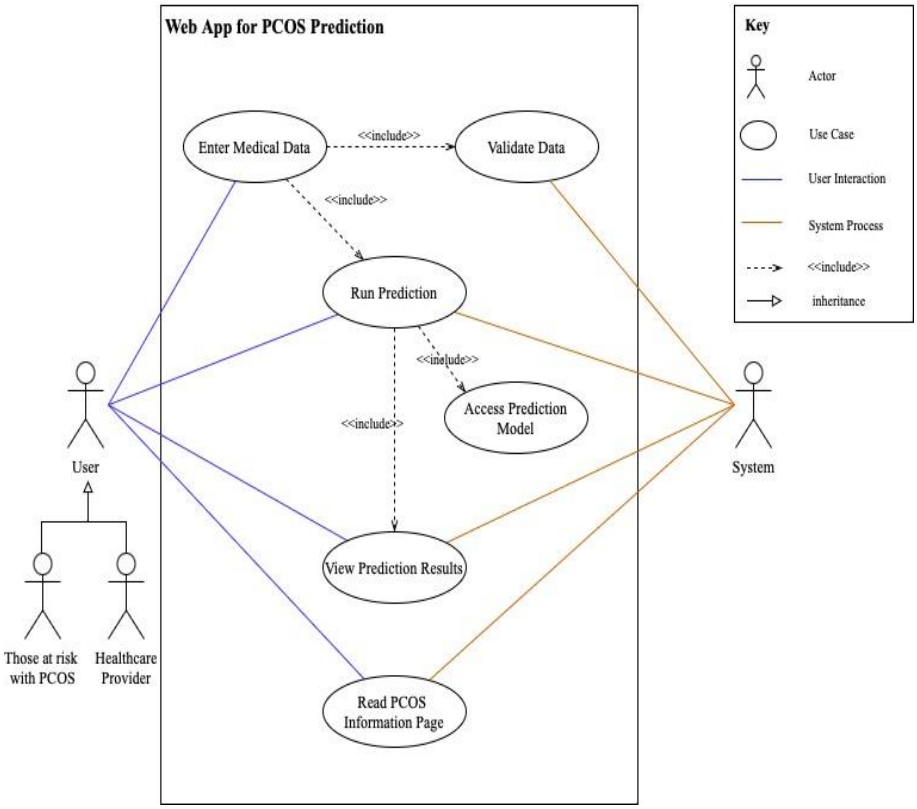


Figure 4: Use Case Diagram (Self composed)

5.2.3 Activity Diagram

This diagram depicts the flow of of processes through different steps in the system.

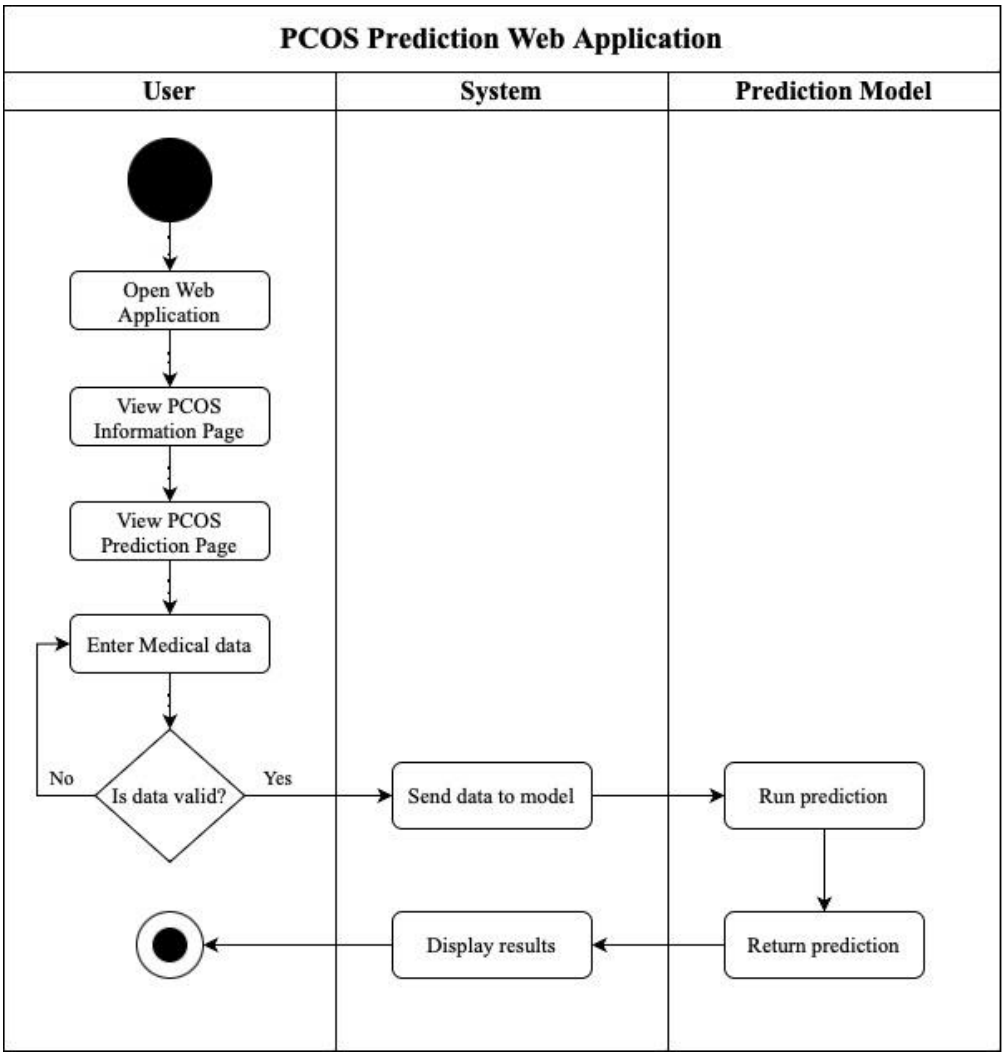


Figure 5: Activity Diagram (Self composed)

5.2.4 Sequence Diagram

This diagram shows how the system objects interact with each other.

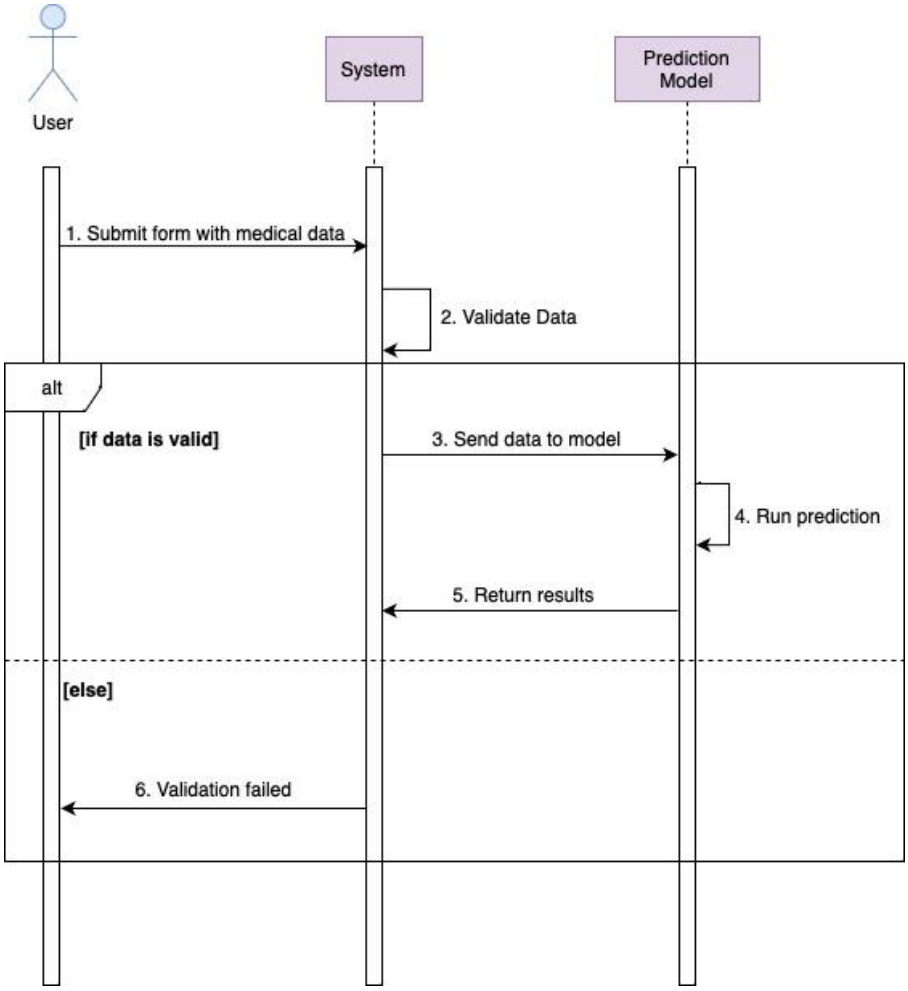


Figure 6: Sequence Diagram (Self composed)

6. Requirements Elicitation

6.1 Stakeholder Analysis

6.1.1 Stakeholder Descriptions

Below are the different types of stakeholders associated with the PCOS Web Application.

Stakeholder	Description
User	Individuals who may show symptoms of PCOS or have a family history of the condition.
Healthcare Providers	Physicians, gynecologists, and healthcare professionals diagnosing and treating PCOS.
Researchers	Researchers from Academia and businesses examining PCOS and its effects.
Healthcare Managers	Those in charge of managing medical facilities and services.
Regulatory Bodies	Organizations that establish rules and regulations for healthcare applications.
Data Privacy Regulators	Authorities responsible for protecting personal data and ensuring privacy laws are followed.
Competitor Systems	Other platforms or studies offering PCOS prediction or management tools.
Women's Health Advocates	Groups and individuals working to raise awareness and educate the public about PCOS and other women's health.
Sponsors	Those who may have interest in funding the development of the application.

Table 8: Stakeholder Descriptions

6.1.2 Onion Model

The Onion Model shows a system's design by arranging many stakeholder layers and their interactions.

The core of the model is the **System Layer**, which is made up of the essential components responsible for the system's primary function.
Direct user communication with the system takes place at the surrounding **Containing System Layer.**
Beyond this is the **Wider Environment Layer**, which contains external entities connected to the system.

This tiered structure shown below reflects the levels of involvement and influence that stakeholders have over the system.

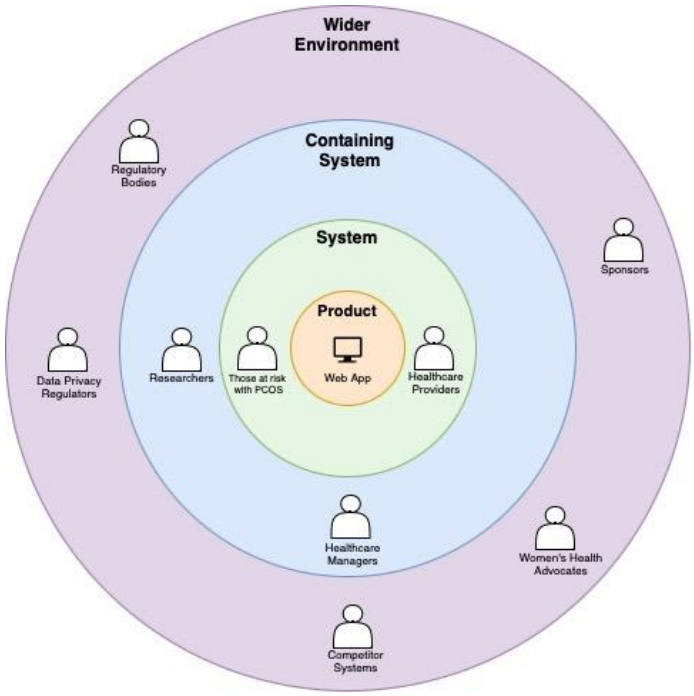


Figure 7: Onion Model (Self composed)

6.2 Functional and Non-functional Requirements

6.2.1 Functional Requirements

Requirements list		Description
FR1	Data Input	Users are able to input health data and symptoms.
FR2	Prediction Model	The system provides predictions based on input data.
FR3	Results Display	Results and recommendations are displayed to users.
FR4	Information Resources	The app provides educational resources on PCOS.

Table 9: Functional Requirements

6.2.2 Non-functional Requirements

Requirements list		Description
NFR1	Performance	The app is efficient and predictions load in an instant.
NFR2	Usability	The app has an intuitive user-friendly interface.
NFR3	Security	User data is not be stored or shared with any third parties.
NFR4	Compatibility	The app is compatible with major web browsers.

Table 10: Non-functional Requirements

6.3 Approaches

Literature Review:

A thorough review of recent publications, websites, and applications relating to PCOS prediction and medical technology was a part of the literature review. This made it easier to find similarities, limitations, and user expectations in web apps for healthcare. The goal was to learn more about important features including feature selection, model selection, prediction accuracy.

Dataset Analysis:

This project will include a detailed analysis of a publicly available PCOS dataset from Kaggle, which comprises anonymized data collected from 10 hospitals across Kerala, India (Kottarathil, 2021). The dataset provides key variables essential for predicting PCOS and related infertility issues, encompassing physical and clinical parameters associated with the condition, with 44 features and 541 rows of data.

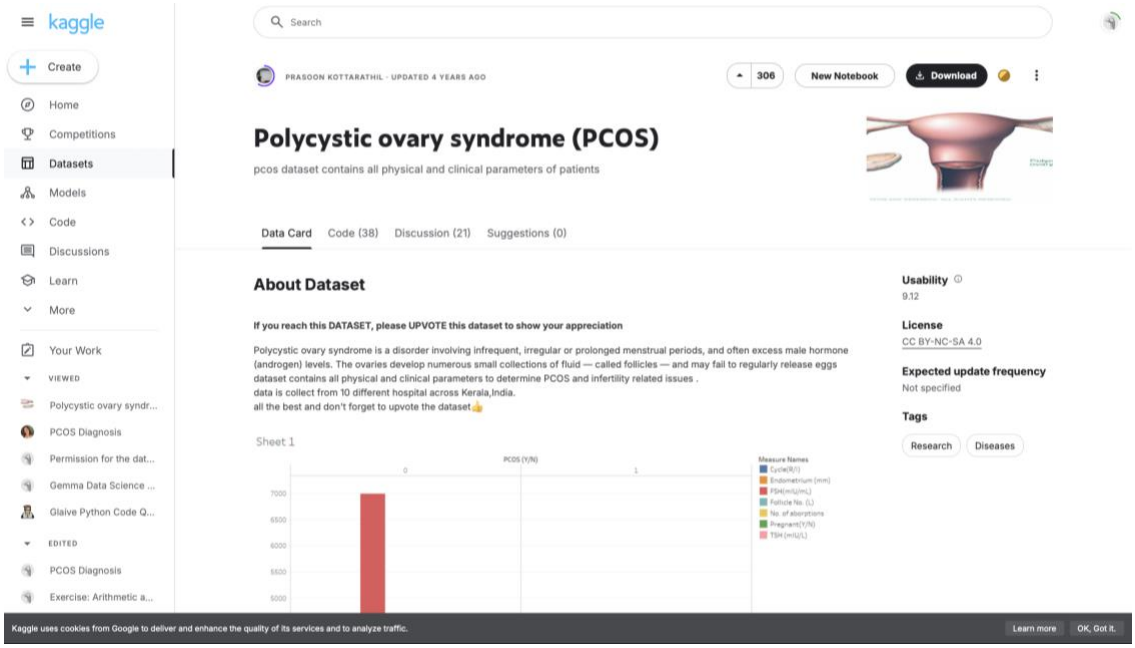
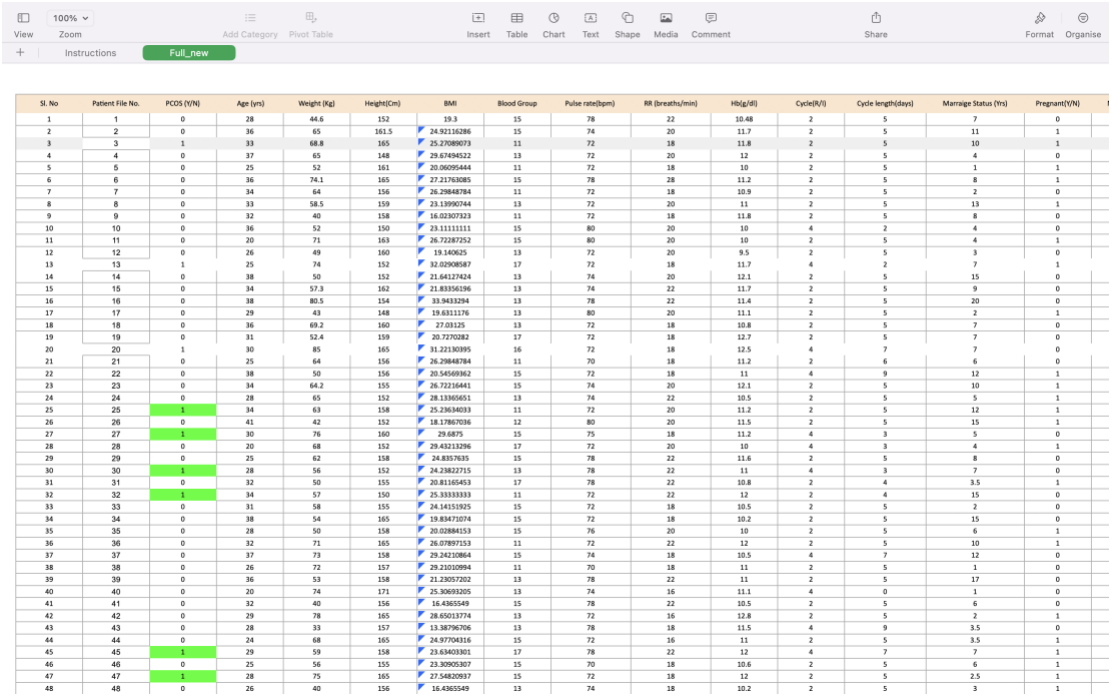


Figure 8: Dataset selected -1

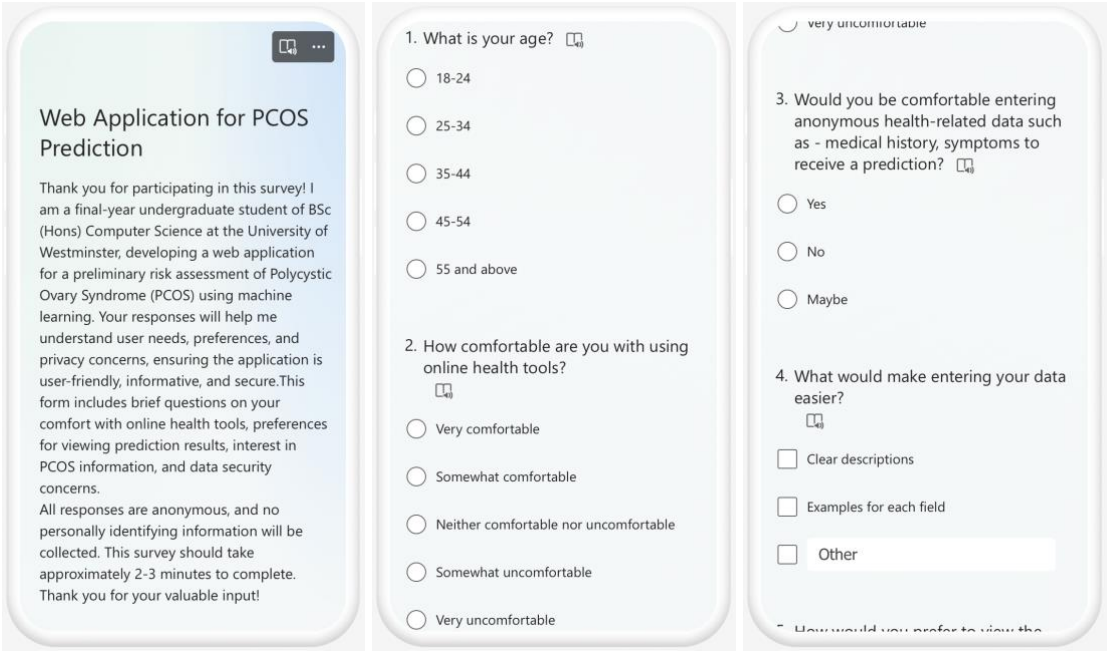


Sl. No	Patient File No.	PCOS (Y/N)	Age (yrs)	Weight (Kg)	Height(Cm)	BMI	Blood Group	Pulse rate(bpm)	RR (breaths/min)	Hb(g/dl)	Cycle(R/I)	Cycle length(days)	Marraige Status (Yrs)	Pregnant(Y/N)
1	1	0	28	44.6	152	19.3	15	78	22	10.48	2	5	7	0
2	2	0	36	65	161.5	24.92316286	15	74	20	11.7	2	5	11	1
3	3	1	33	68.8	165	25.27080073	11	72	18	11.8	2	5	10	1
4	4	0	37	65	148	29.67494522	13	72	20	12	2	5	4	0
5	5	0	25	52	161	20.06095444	11	72	18	10	2	5	1	1
6	6	0	36	74.1	165	27.21763085	15	78	28	11.2	2	5	8	1
7	7	0	34	64	156	26.29848784	11	72	18	10.9	2	5	2	0
8	8	0	33	58.5	159	23.13990744	13	72	20	11	2	5	13	1
9	9	0	32	40	158	16.02307323	11	72	18	11.8	2	5	8	0
10	10	0	36	52	150	23.11111111	15	80	20	10	4	2	4	0
11	11	0	20	71	163	26.72287252	15	80	20	10	2	5	4	1
12	12	0	26	49	160	19.140625	13	72	20	9.5	2	5	3	0
13	13	1	25	74	152	32.02908587	17	72	18	11.7	4	2	7	1
14	14	0	38	50	152	21.64127424	13	74	20	12.1	2	5	15	0
15	15	0	34	57.3	162	21.83356196	13	74	22	11.7	2	5	9	0
16	16	0	38	80.5	154	33.9433294	13	78	22	11.4	2	5	20	0
17	17	0	29	43	148	19.6311176	13	80	20	11.1	2	5	2	1
18	18	0	36	69.2	160	27.03125	13	72	18	10.8	2	5	7	0
19	19	0	31	52.4	159	20.7270282	17	72	18	12.7	2	5	7	0
20	20	1	30	85	165	31.22130395	16	72	18	12.5	4	7	7	0
21	21	0	25	64	156	26.29848784	11	70	18	11.2	2	6	6	0
22	22	0	38	50	156	20.54569362	15	72	18	11	4	9	12	1
23	23	0	34	64.2	155	26.72216441	15	74	20	12.1	2	5	10	1
24	24	0	28	65	152	28.13365651	13	74	22	10.5	2	5	5	1
25	25	1	34	63	158	25.23634033	11	72	20	11.2	2	5	12	1
26	26	0	41	42	152	18.17867036	12	80	20	11.5	2	5	15	1
27	27	1	30	76	160	29.6875	15	75	18	11.2	4	3	5	0
28	28	0	20	68	152	29.43213296	17	72	20	10	4	3	4	1
29	29	0	25	62	158	24.8357635	15	78	22	11.6	2	5	8	0
30	30	1	28	56	152	24.23822715	13	78	22	11	4	3	7	0
31	31	0	32	50	155	20.81165453	17	78	22	10.8	2	4	3.5	1
32	32	1	34	57	150	25.33333333	11	72	22	12	2	4	15	0
33	33	0	31	58	155	24.14151925	15	72	18	10.5	2	5	2	0
34	34	0	38	54	165	19.83471074	15	72	18	10.2	2	5	15	0
35	35	0	28	50	158	20.02884153	15	76	20	10	2	5	6	1
36	36	0	32	71	165	26.07897153	11	72	22	12	2	5	10	1
37	37	0	37	73	158	29.24210864	15	74	18	10.5	4	7	12	0
38	38	0	26	72	157	29.21010994	11	70	18	11	2	5	1	0
39	39	0	36	53	158	21.23057202	13	78	22	11	2	5	17	0
40	40	0	20	74	171	25.30693205	13	74	16	11.1	4	0	1	0
41	41	0	32	40	156	16.4365549	15	78	22	10.5	2	5	6	0
42	42	0	29	78	165	28.65013774	13	72	16	12.8	2	5	2	1
43	43	0	28	33	157	13.38796706	13	78	18	11.5	4	9	3.5	0
44	44	0	24	68	165	24.97704316	15	72	16	11	2	5	3.5	1
45	45	1	29	59	158	23.63403301	17	78	22	12	4	7	7	1
46	46	0	25	56	155	23.30905307	15	70	18	10.6	2	5	6	1
47	47	1	28	75	165	27.54820937	15	72	18	12	2	5	2.5	1
48	48	0	26	40	156	16.4365549	13	74	18	10.2	2	5	3	1

Figure 9: Dataset selected -2

Questionaries:

Surveys will be distributed to potential users to gather their thoughts on preferred features, usability, and concerns related to PCOS prediction applications. The purpose of these questionnaires is to ensure the application aligns with user needs and expectations.



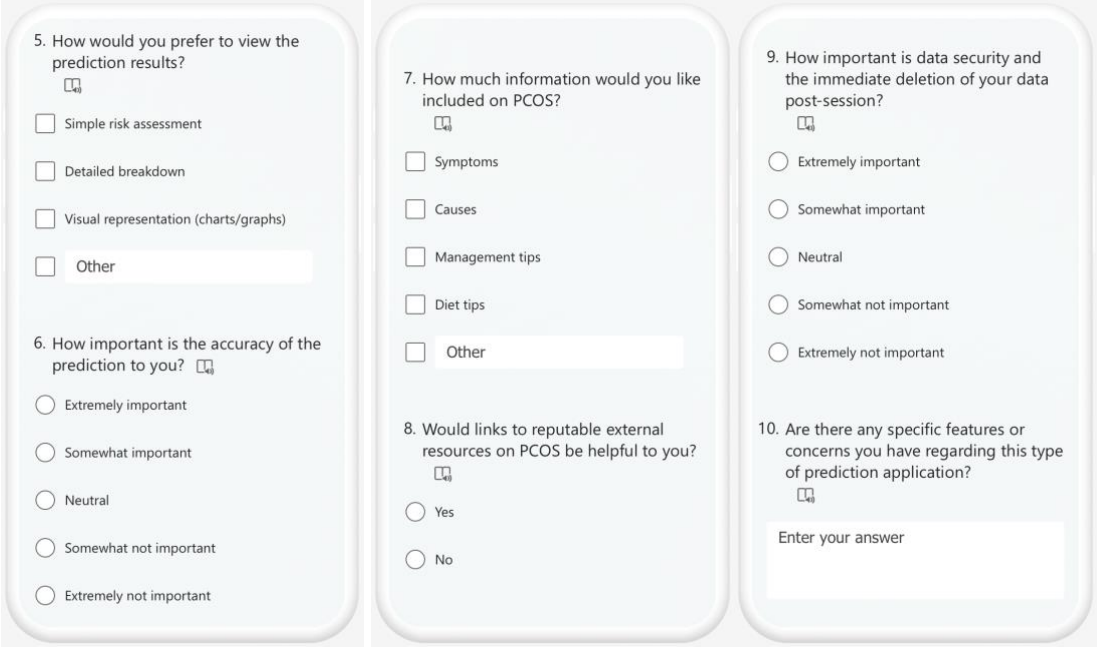


Figure 10: Questionnaire

Brainstorming:

The purpose was to explore and come up with innovative solutions and consider diverse ideas that would enhance the application's development.

6.4 Techniques

Data Analysis

It will be used to examine dataset attributes, clean the dataset, perform feature selection, and understand variable relationships, to make sure essential information is captured in the application's data input form.

System Requirements Analysis

Identified functional and non-functional requirements based on insights from the literature review and dataset, focusing on user-friendly design, predictive accuracy, and educational content.

User Interface (UI) Design Prototyping

Create wireframes and prototypes to visualize the user interface and interaction flow, allowing early feedback through user testing.

Agile Methodology

An iterative development process allows regular feedback and adjustments based on user testing and stakeholder input, ensuring alignment with user needs and expectations.

7. Time Schedule

7.1 Gantt Chart Overview

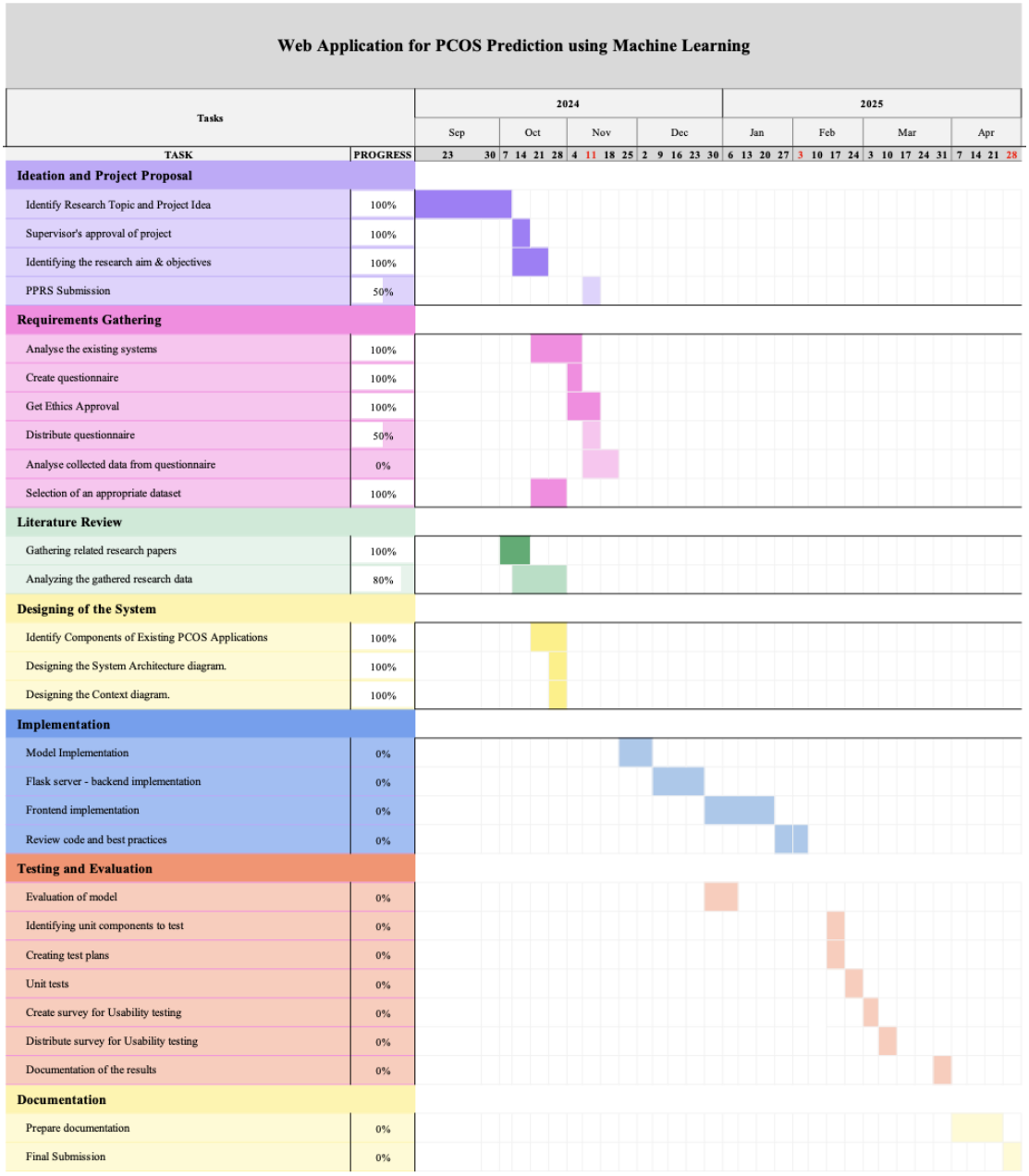


Figure 11: Gantt Chart (Self composed)

8. References

Adla, Y. A. et al. (2021). Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques. 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME). Available from http://dx.doi.org/10.1109/ICABME53305.2021.9604905 [Accessed 6 October 2024].

Gandhi, D., Patel, B. and Dave, N. (2024). PCOS Detection using Machine Learning Algorithms. Indian Journal of Natural Science, 14 (81). Available from https://www.researchgate.net/publication/381456665_PCOS_Detection_using_Machine_Learning_Algorithms [Accessed 8 October 2024].

Hdaib, D. et al. (2022). Detection of Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms. 5th International Conference on Engineering Technology and its Applications 2022. Available from https://doi.org/10.1109/IICETA54559.2022.9888677 [Accessed 6 October 2024].

Kottarathil, P. (2021). Polycystic ovary syndrome (PCOS). Available from https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos/data [Accessed 6 October 2024].

NHS. (2022). Polycystic ovary syndrome. Available from https://www.nhs.uk/conditions/polycystic-ovary-syndrome-pcos/ [Accessed 6 October 2024].

Priyadharshini, M. et al. (2024). PCOS Disease Prediction Using Machine Learning Algorithms. International Research Journal on Advanced Engineering Hub (IRJAEH), 2, 651—655. Available from http://dx.doi.org/10.47392/IRJAEH.2024.0094 [Accessed 6 October 2024].

Radhakrishnan, R. et al. (2023). Computerized Diagnosis of Polycystic Ovary Syndrome Using Machine Learning and Swarm Intelligence Techniques. Available from https://doi.org/10.21203/rs.3.rs-2027767/v2 [Accessed 6 October 2024].

Tiwari, S. et al. (2022). SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning. Expert Systems with Applications 203, 1. Available from http://dx.doi.org/10.1016/j.eswa.2022.117592 [Accessed 6 October 2024].

9. Ethics Form

UNIVERSITY OF WESTMINSTER | SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Final Year Project

Research Project Submission Form

Section 0: Applicant's info

Applicant's Full Name:	Rachel Naadia Cooray
Applicant's University ID number:	19564441
University Email Address:	w1956444@my.westminster.ac.uk
Select your level of study:	Undergraduate

Section 1: Project Info

1.1. Supervisor's Full Name:	Dr Malarvizhi Kaniappan Chinnathai
1.2. Supervisor's Email Address:	M.Kaniappanchinnathai@westminster.ac.uk
1.3. Project Title:	Web Application for Diagnosing PCOS Patients Using Machine Learning
1.4. Please provide a description of the background with references to relevant literature (250 words):	Polycystic Ovary Syndrome (PCOS) is a common hormonal disorder affecting women of reproductive age. The disorder is characterized by irregular menstrual cycles, high levels of male hormones known as androgens, and the presence of cysts in the ovaries, which can lead to a variety of symptoms such as acne, excessive hair growth, obesity, and difficulty in conceiving (Hdaib et al., 2022). The exact cause of PCOS is not fully understood, though genetic factors and lifestyle influences, such as diet and obesity, are believed to contribute (Gandhi et al., 2023). Due to its diverse symptoms and the variation in individual presentations, PCOS is often underdiagnosed, with many women remaining undiagnosed until later in life (Cicek and Kucukakcali, 2024). Early detection of PCOS is crucial as the disorder is associated with significant long- term health risks, including cardiovascular diseases, type 2 diabetes, endometrial cancer, and mental health issues (Gandhi et al., 2023). Traditional diagnostic methods rely on clinical symptoms and imaging techniques like ultrasound, but they are not always sufficient due to overlapping symptoms with other conditions (Hdaib et al., 2022).

Figure 12: Ethics Form -1

This has led to an increased interest in utilizing machine learning (ML) models for more accurate, efficient, and early detection of PCOS. Machine learning algorithms, such as Random Forest, Support Vector Machines, and Decision Trees, have been successfully employed in detecting PCOS based on patient data, offering high accuracy and predictive power (Radhakrishnan et al., 2023).

References

Cicek, I. B. and Kucukakcali, Z (2024). Global interpretation of machine learning models in predicting polycystic ovary syndrome with the explainable artificial intelligence method SHAP. Medicine Science 2024, 13(3), 568-74. Available from 10.5455/ medscience.2023.12.228 [Accessed 6 October 2024].

Gandhi, D., Patel, B. and Dave, N. (2023). PCOS Detection using Machine Learning Algorithms. Indian Journal of Natural Science, 14 (81). Available from https:// www.researchgate.net/ publication/381456665_PCOS_Detection_using_Machine_ Learning_Algorithms [Accessed 8 October 2024].

Hdaib, D. et al. (2022). Detection of Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms. 5th International Conference on Engineering Technology and its Applications 2022. Available from https:// doi.org/10.1109/ IICETA54559.2022.9888677 [Accessed 6 October 2024].

Radhakrishnan, R. et al. (2023). Computerized Diagnosis of Polycystic Ovary Syndrome Using Machine Learning and Swarm Intelligence Techniques. Available from https:// doi.org/10.21203/ rs.3.rs-2027767/ v2 [Accessed 6 October 2024].

1.5. Please provide a brief description and the aims of your study (250 words):	The objective of this study is to develop and assess a machine learning- based system designed to detect Polycystic Ovary Syndrome (PCOS) at an early stage, using patient data. This research aims to explore how machine learning models can enhance the accuracy and efficiency of PCOS diagnosis. By using a dataset that includes various medical and hormonal features, the model will be trained to recognize patterns linked to PCOS, such as hormonal imbalances, menstrual irregularities, and physical symptoms like acne and excessive hair growth. Also, the study will compare the performance of different machine learning algorithms, including Random Forest, Support Vector Machines, and Decision Trees, to identify the most effective model for diagnosing PCOS. The evaluation will focus on key metrics such as accuracy, precision, recall, and F1-score, for the model's reliability and robustness. Ultimately, the aim is to contribute to healthcare technology by providing a tool that supports the timely and accurate identification of PCOS.
1.6. Please outline the design and methodology of your	The project aims to develop a machine learning based

Figure 13: Ethics Form -2

study and details of any invasive or intrusive procedures (400 words):	web application for the early detection of Polycystic Ovary Syndrome (PCOS) using patient data. The study will utilize an existing, publicly available dataset, from Kaggle (https:// www.kaggle.com/ datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos/data), which includes medical and hormonal data of women, such as age, weight, insulin levels, LH/ FSH ratios, and symptoms like acne, hair growth, and irregular menstruation. This dataset contains both categorical and continuous features, allowing for the training of multiple machine learning models. Data cleaning will be performed to handle missing values, outliers, and inconsistent entries. The project will explore several machine learning algorithms to determine which provides the most accurate and reliable predictions for PCOS diagnosis. The models will be evaluated using metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques will be used to ensure that the model performs well on unseen data. Meanwhile, the frontend web pages will be developed and the machine learning model will be integrated to it. No invasive or intrusive procedures are involved in this study. All data used for model training and evaluation will be anonymized, and the use of patient data will adhere strictly to ethical guidelines and data protection regulations.
1.7. Project Start Date:	2024-09-23
1.8. End Date of Work:	2025-04-28

Section 2: External Factors

2.1. Does your research include funding from an external organisation and/or external collaborator/s or co-Investigator/s?:	NO
2.2. Are you seeking ethical approval from the Health Research Authority (HRA)?:	NO
2.3. Are you seeking University sponsorship (as defined by Health Research Authority)?:	NO
2.4. Are you seeking ethical approval from any other external organisation (which is not the Health Research Authority)?:	NO
2.5. Have you been asked by an external organisation to produce evidence of ethical approval for your research?:	NO

Section 3: Participants

3.1. Does this research proposal (as proposed to Research Ethics Committee in its current status) include Research Participants (humans and/ or animals, either deceased or alive)?:	YES
3.3. Human participants in Health and Social Care settings?:	NO

Figure 14: Ethics Form -3

3.4. Human participants who may be deemed vulnerable due to their setting(s)?:	NO
3.5. Expectant or new mothers?:	NO
3.6. Refugees or asylum seekers or recent migrants?:	NO
3.7. Minors (under the age of 18 years old)?:	NO
3.8. Participants in custody (e.g. prisoners or arrestees)?:	NO
3.9. Participants who may potentially fall under the remit of the Mental Capacity Act?:	NO
3.10. Are animals (or animal tissue) involved?:	NO

Section 4: Risk of harm

4.1. Will any pain or more than mild discomfort result from the study?:	NO
4.2. Could the study induce any psychological stress or anxiety or cause harm or negative consequences beyond the risks encountered in normal life?:	NO
4.3. Will the study involve prolonged or repetitive physical or psychological testing of human participants that may put someone at risk, e.g. use of treadmill?:	NO
4.4. Will the study involve raising sensitive topics (e.g. sexual activity, drug use, revelation of medical history, bereavement, illegal activities, etc.)?:	NO
4.5. Does your work involve relevant material, defined by the Human Tissue Act as material other than gametes, which consists of, or includes, human cells. In the Human Tissue Act, references to relevant material from a human body do not include: (a) embryos outside the human body, or (b) hair and nail from the body of a living person. Work falling under the Human Tissue Authority.:	NO
4.6. Will DNA samples be taken from human participants?:	NO
4.7. Does your study raise any issues of personal safety for you or other researchers or participants involved in the project (especially relevant if taking place outside working hours or off-site e.g. not on University premises)?:	NO
4.8. Does your study involve deliberately misleading the participants (e.g. deception, covert observation)?:	NO
4.9. Does your work involve administration of a food or non-food substance of a different type from or in abnormally higher or lower amounts than normal or one that is known to cause allergic reaction(s) or potential psychological stress?:	NO
4.10. Does your study involve issues relating to personal and/or sensitive data?:	NO
4.11. Does your research involve any 'security sensitive material? See Universities UK Oversight of Security Sensitive Research Material (2019).:	NO
4.12. Does your research ethics proposal include off-site (i.e. not on University premises) research fieldwork and travel involving face to face interactions?:	NO

Section 5: Information to participants

Figure 15: Ethics Form -4

5.1. Will you provide participants with a Participant Information Sheet prior to obtaining informed consent?:	YES
5.2. Will you describe the procedures to participants in advance, so that they are informed about what to expect?:	YES
5.3. Will you obtain informed consent for participation (normally written)?:	YES
5.4. Will you tell participants that they may withdraw from the research at any time and for any reason?:	YES
5.5. Will you give participants the option of omitting questions they do not want to answer?:	YES
5.6. Will you tell participants that their data will be treated as confidential and that, if published, it will not be identifiable as theirs?:	YES
5.7. Will you offer feedback to participants at the end of their participation, upon request (e.g. give them a brief explanation of the study and its outcomes)?:	YES

The applicant confirmed and agreed the following statements:

- The information I have given on this form is true, complete and to the best of my knowledge correct.
- They have read the University's Code of Practice Governing the Ethical Conduct of Research.
- The information provided on this form is subject to the Data Protection Act 2018, General Data Protection Regulation (GDPR) 2018 and the Freedom of Information Act 2000.
- This form may be disclosed as a result of a GDPR Subject Access Request.
- This form may be disclosed as a result of a request for information under the Freedom of Information Act 2000.
- They must ensure that any subjects selected for study are made aware of their rights and our obligations under the Data Protection Act 2018 and General Data Protection Regulation (GDPR) 2018.
- They must ensure that sponsors are made aware that the University of Westminster is subject to the Freedom of Information Act 2000.

Applicant's signature

Date: 8th November 2024

Supervisor's signature

Malarvizhi Kaniappan Chinnathai

Date: 11/11/2024

This form was completed and submitted by the applicant [Rachel Naadia Cooray] on Friday, 8 Nov 2024

Figure 16: Ethics Form -5