Machine Learning-Enhanced Detection of Polycystic Ovary Syndrome

HAJAR DYA

Supervised By : Mr .AZIZ KHAMJANE

Abdelmalek Essaâdi University, National School of Applied Sciences Al-Hoceima

Hajar.dya@etu.uae.ac.ma

Abstract

The hormonal complexity of Polycystic Ovary Syndrome presents a significant challenge in women's health, affecting between 5 and 10 percent of reproductive-age females globally. This endocrine disorder manifests through a characteristic triad: androgens exceeding normal levels, disrupted menstrual patterns, and metabolic alterations. The stakes of delayed diagnosis are considerable, potentially leading to compromised fertility, heightened cardiovascular risk, and increased diabetes susceptibility.

I- Introduction :

The hormonal complexity of Polycystic Ovary Syndrome presents a significant challenge in women's health, affecting between 5 and 10 percent of reproductive-age females globally. This endocrine disorder manifests through a characteristic triad: androgens exceeding normal levels, disrupted menstrual patterns, and metabolic alterations. The stakes of delayed diagnosis are considerable, potentially leading to compromised fertility, heightened cardiovascular risk, and increased diabetes susceptibility.

Our investigation leverages contemporary machine learning approaches to enhance diagnostic precision for PCOS. Through deep analysis of clinical investigations and symptom correlations, we aim to establish more reliable detection methods. The research framework incorporates extensive data points spanning hormonal profiles, metabolic markers, and clinical presentations, utilizing sophisticated machine learning algorithms to optimize diagnostic capabilities.

Central to our methodology is the pursuit of clinically meaningful results, with particular emphasis on three key metrics: diagnostic accuracy, sensitivity in detection, and specificity of identification. This research endeavors to transform current diagnostic paradigms by providing healthcare practitioners with evidence-based, technologically advanced tools for both early detection and individualized treatment strategies in PCOS care management.

II- Previous Studies :

The application of machine learning (ML) and deep learning (DL) in detecting Polycystic Ovary Syndrome (PCOS) has been a focal point in recent research, showcasing innovative methodologies and leveraging diverse data sources. In the study titled **"PCONet: A Convolutional Neural Network Architecture to Detect Polycystic Ovary Syndrome (PCOS) from Ovarian Ultrasound Images,"** researchers developed a convolutional neural network (CNN) specifically for analyzing ovarian ultrasound images. The model achieved an impressive accuracy of

98.12%, demonstrating its capability in identifying polycystic features with high precision. This work underscores the potential of CNNs in automating the interpretation of complex medical imagery, reducing manual errors, and enhancing diagnostic efficiency.
A distinct approach was explored in **"Deep Learning Algorithm for Automated Detection of Polycystic Ovary Syndrome Using Scleral Images,"** where scleral imaging was used as a novel biomarker for PCOS detection. By employing a ResNet-based framework for feature extraction combined with a multi-instance classification model, this study achieved an AUC of 0.979 and a classification accuracy of 92.9%. This innovative technique provided a non-invasive and practical alternative for identifying PCOS, broadening the scope of diagnostic tools in clinical practice. **"Polycystic Ovary Syndrome Detection Machine Learning Model"** focused on traditional ML approaches to diagnose PCOS using clinical and symptomatic data. The study evaluated several algorithms, including logistic regression, random forest, and XGBoost, emphasizing their predictive power and interpretability. By prioritizing model explainability, the study aimed to build trust in automated diagnostic systems and highlight the integration of patient-centric data for reliable predictions.Collectively, these studies reveal the growing role of AI-driven techniques in PCOS detection. From deep learning models analyzing intricate imaging data to machine learning algorithms processing clinical features, these approaches illustrate the potential of hybrid and multimodal strategies in improving diagnostic accuracy and accessibility.

III- Material and Methodology :

The study methodology followed a systematic machine learning approach. Initially, we acquired the dataset and performed feature selection by eliminating non-contributory columns to enhance model efficiency. The preprocessed dataset was then partitioned into training and testing sets to ensure unbiased model evaluation.

In the first phase of modeling, we implemented and compared six different machine learning algorithms:

- **Logistic Regression (LR)**
- **Decision Tree (DT)**
- **Random Forest (RF)**
- **Support Vector Machine (SVM)**
- **K-Nearest Neighbors (KNN)**
- **Naive Bayes (NB)**

Comparative analysis of these models revealed that Logistic Regression demonstrated superior performance metrics among all tested algorithms.
Subsequently, we conducted a second phase of analysis focused on patient accessibility. We refined our feature set to include only patient-reportable variables, excluding clinical measurements that would require medical intervention to obtain. This refined dataset was then subjected to the same suite of machine learning algorithms. In this second evaluation phase, the Support Vector Machine (SVM) classifier emerged as the optimal model, exhibiting the

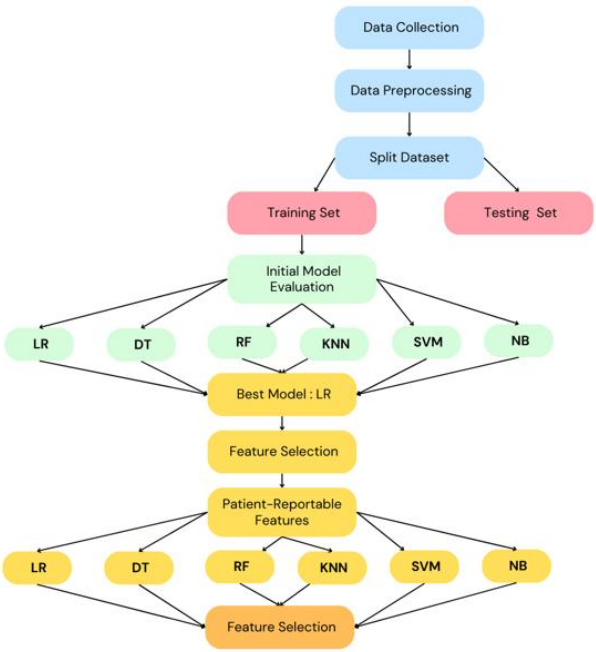

Figure 1 : Project Workflow

best performance characteristics when utilizing solely patient-reportable features.
This two-phase approach enabled us to identify not only the best-performing model using all available data but also the most effective algorithm for a more practical, patient-centered implementation.

Database Description :

The study utilized the Polycystic Ovary Syndrome (PCOS) dataset obtained from Kaggle [26]. The dataset comprises 541 patient records with 41 distinct attributes. The class distribution reveals 178 positive cases (PCOS) and 363 negative cases (non-PCOS), indicating an imbalanced class distribution. The complete dataset was constructed by merging two complementary files: "PCOS_infertility" and "PCOS_data_without_infertility," followed by the elimination of redundant columns to ensure data integrity. Table 1 presents a comprehensive overview of the dataset features and their characteristics.

Table 1: Table of features

Feature Name	**Abb**	**Description**
Patient File No.		Patient file number (unique identifier)
Polycystic Ovary Syndrome	PCOS	Class label
Age		
Weight	WEIGHT	Patient's weight in KG
Height	HEIGHT	Patient's height in CM
Body Mass Index	BMI	Body mass index of the patient (height/weight)
Blood Group	BG	Patients belong to which blood group (A+, A−, B+, B−, O+, O−, AB+, AB−)
Pulse Rate	PR	Heartbeat per minute
Respiration Rates	RR	Respiration rates per minute
Hemoglobin	HB	Number of red blood cells in patient's body
Cycle	CYCLE	Length of the menstrual cycle
Cycle Length	CL	Number of days of a cycle
Marriage Status	MS	Number of years since marriage
Pregnant	P	Pregnant status
No. of Abortions	AB	No. of abortions
I Beta-HCG	BETA_I	Amount of human chorionic gonadotropin
Beta Healthy	BETA_II	Beta HCG level is indication of 100 mIU/ml
Singleton Pregnancy		about 16 days after ovulation,
Follicle-Stimulating Hormone	FSH	Attributes ranging from 0.3 to 10.0 mIU/mL indicate if are still menstruating or have undergone menopause
Luteinizing Hormone	LH	Chemical agitator that stimulates the reproductive system
Follicle-Stimulating Hormone/ Luteinizing Hormone	FSH/LH	Ratio of FSH and LH
Hip Size	HIP	Size of hip in inches
Waist Size	WAIST	Size of waist in inches
Waist-Hip Ratio	HIP_RATIO	Waist size proportion to hip
Thyroid-Stimulating Hormone	TSH	Amount of TSH in the blood
Anti-Mullerian Hormone	AMH	Plays a key role in developing a baby's sex organs while in the womb
Prolactin levels	PRL	Prolactin levels in women's bodies
Vitamin D	VIT_D3	Vitamin D levels
Progesterone Levels	PRG	Progesterone levels
Random Blood Sugar	RBS	Value of random blood sugar (RBS) test
Weight Gain	WG	Test to check if the patient gains weight
Hair Growth	HG	Test to check if a patient has hair growth
Skin Darkening	SD	Test to check the appearance of darkness in skin
Hair Loss	HL	Test to check hair loss
Pimples	PIMPLES	Pimple issues
Fast Food	FF	Check if fast food part of the diet
Reg.Exercise	RE	Check if patient exercises on a regular basis
Blood Pressure Systolic	BP_ SYSTOLIC:	Amount of pressure in the arteries when the heart is contracting
Blood Pressure Diastolic	BP_ Diastolic	Amount of pressure in the arteries while the heart is resting in between heart beats
Follicle No.	FN	Follicle number in the left side

In the preprocessing phase of our medical dataset, we implemented a comprehensive numerical encoding system for blood type classification to optimize machine learning analysis. This transformation was essential as most algorithms require numerical inputs rather than categorical data. The encoding scheme was carefully designed to reflect the biological complexity and characteristics of different blood types.

The blood groups were systematically encoded using a numerical sequence from 11 to 18, with the values assigned in a pattern that considers both the ABO system and the Rh factor.

A+ = 11 A- = 12

B+ = 13 B- = 14

O+ =15 O- = 16

AB+ =17 AB- = 18

2. Pre-Processing :

2.1 Missing Values Treatment :

The dataset requires preprocessing to address quality issues common in medical data collection. Two main challenges are addressed:

Missing Values Treatment:

- Features with >30% missing data are removed (BMI, FSH/LH, Waist:Hip Ratio)
- Administrative columns (Sl. No, Patient File No.) are eliminated
- For features with <30% missing values (Marriage Status, II beta-HCG, AMH, Fast food), the median value is used as a replacement

This approach balances data preservation with reliability. The 30% threshold represents a practical cutoff point - beyond this, data imputation risks introducing significant bias. Median imputation was selected for its robustness to outliers compared to mean imputation, particularly important for medical measurements which often show skewed distributions.

This preprocessing ensures the dataset's suitability for machine learning models while maintaining clinical relevance and statistical validity.

2.2 Converting Columns to Numerical Values :

Transforming categorical variables into numerical values is a critical step in machine learning, as most algorithms are designed to process numerical data exclusively. This conversion ensures that the inherent relationships within the data are preserved, facilitating effective computational analysis and making the dataset suitable for advanced tasks such as classification.

2.3 Exploratory Data Analysis (EDA):

a. Analysis of Correlation Using Graphs:

Based on our graphical analysis, we have identified significant correlations between age groups and the prevalence of PCOS (Polycystic Ovary Syndrome):

- Women aged 26–35 years are the most affected by PCOS, followed by those aged 18–25 years.
- Women aged 45 years and older are the least affected, indicating that age plays

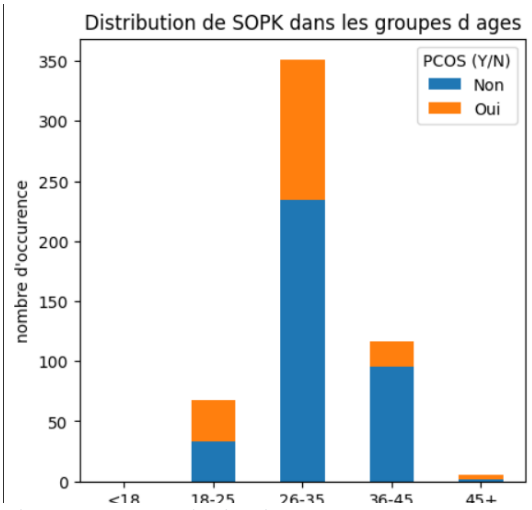

Figure 2 : PCOS Distribution Across Age Groups

a crucial role in the prevalence of the condition.

b. Numerical Data Analysis for Women Diagnosed with PCOS :

To further explore correlations, we analyzed numerical variables that are not boolean. The findings reveal key patterns that enhance our understanding of PCOS and its associated factors:

1. Body Mass Index (BMI):

A BMI of 25, observed with the highest frequency, highlights a strong correlation between obesity and the occurrence of PCOS.

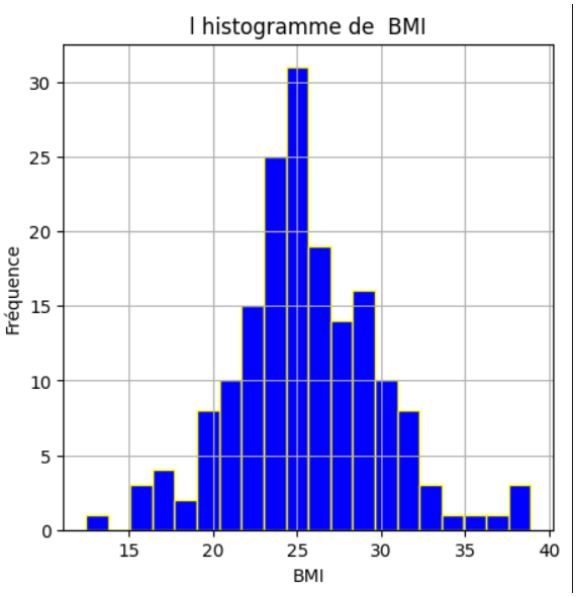

Figure 4: BMI Distribution Among Women with Polycystic Ovary Syndrome (PCOS)

2. Hemoglobin Levels:

Hemoglobin levels between 10 and 11 g/dL are lower than the normal range, indicating that anemia is common among women with PCOS. This condition often leads to fatigue and reduced energy levels.

3. Waist-to-Hip Ratio (WHR):

A WHR greater than 0.95, surpassing normal thresholds, is linked to abdominal fat accumulation, which increases the risk of diabetes.

4. Anti-Müllerian Hormone (AMH):

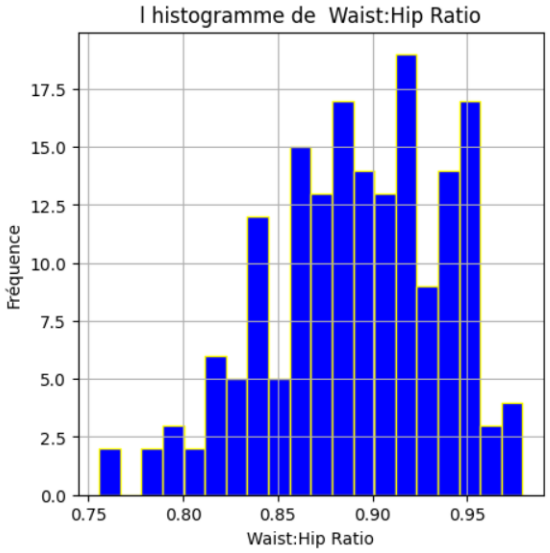

Figure 3: HIP Ratio Distribution Among Women with Polycystic Ovary Syndrome (PCOS)

Under normal conditions, AMH levels range from 1 to 4 ng/mL. Elevated levels are a typical biomarker of PCOS.

5. Prolactin Levels:

Normal prolactin levels range from 4.8 to 23.3 ng/mL. Abnormal values suggest hormonal imbalances frequently associated with PCOS.

6. Random Blood Sugar (RBS):

Normal RBS levels fall between 70 and 140 mg/dL. Deviations from this range could indicate glucose metabolism irregularities, which are common in PCOS patients.

7. Endometrial Thickness:

An endometrial thickness of 8 mm is considered normal during the luteal phase. Deviations may indicate reproductive health issues linked to PCOS.

8. Diastolic Blood Pressure (BP):

A diastolic BP of ≥80 mmHg is classified as elevated, suggesting a correlation between hypertension and PCOS.

2.4. Data Splitting :

This involves splitting the dataset into two partitions for training and testing, ensuring the model's effectiveness with new, unseen data. In

our study, we divided the entire dataset into training and testing sets, maintaining an 80%-20% ratio.

4. Approaches :

In our research, we conducted a comprehensive evaluation of machine learning algorithms for PCOS detection through a two-phase methodology. The initial phase assessed multiple algorithms using the complete feature set, while the second phase focused on patient-observable characteristics.

Phase 1: Full Feature Analysis :

The initial phase involved testing a wide range of machine learning models on the cleaned dataset that included all available features. These models were selected based on their general applicability and effectiveness in classification tasks. We evaluated six distinct machine learning classifiers:

1.1- **Logistic Regression (LR)**:

A widely-used linear model for binary classification tasks, well-suited for interpreting relationships between features and outcomes.

$$\textit{Sigmoid Function: } f(x) = \frac{1}{1+e^{-x}}$$

1.2- **Decision Tree**:

A non-linear model that provides interpretability by splitting the data based on feature values to predict the outcome.

1.3- **Random Forest**:

An ensemble method combining multiple decision trees to reduce overfitting and improve accuracy by averaging the predictions of individual trees.

1.4- **K-Nearest Neighbors (KNN)**:

A non-parametric algorithm that classifies data points based on the majority class of their nearest neighbors in feature space.

1.5- **Naïve Bayes**:

A probabilistic model that applies Bayes' theorem, assuming feature independence, to predict the class based on prior probabilities and likelihoods.

1.6- **Support Vector Machine (SVM)**:

A powerful model that finds the optimal hyperplane to separate different classes by maximizing the margin between data points of different classes.

Phase 2: Observable Feature Analysis:

Following clinical considerations, we identified 14 key features readily accessible to patients:

Binary indicators: PCOS status, skin darkening, hair growth/loss, weight gain, cycle regularity, fast food consumption, pimples, exercise habits, pregnancy status

Continuous measurements: weight, waist circumference, hip circumference, cycle duration.

The second phase involved testing a wide range of machine learning models on the cleaned dataset that included all available features. These models were selected based on their general applicability and effectiveness in classification tasks. We evaluated the same machine learning as the first one.

5. Model Evaluation :

In the response dataset **,** Polycystic Ovary Syndrome (PCOS) , diagnosis is validated using four classes:

True Positive (TP): Correctly recorded cases with PCOS.

True Negative (TN): Cases correctly identified as not having PCOS.

False Positive (FP): Incorrectly recorded cases with PCOS.

False Negative (FN): Incorrectly predicted cases without PCOS.

the models are evaluated using four methods: accuracy, precision, recall, and F-score, where TP indicates true positive, TN indicates true negative, FP indicates false positive, and FN indicates false negative.

Accuracy : is a fundamental metric that measures the overall correctness of a model by calculating the ratio of correctly predicted instances to the total instances.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Precision: Correctly predicted autism cases out of all predicted positive cases.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Correctly identified autism cases out of all actual positive cases.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: Harmonic mean of precision and recall, offering a balanced measure.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These measures collectively evaluate the accuracy and effectiveness of the classifiers in predicting ASD.

6. Model Tuning :

Model optimization, or hyperparameter tuning, involves fine-tuning the model's parameters to identify the most suitable configuration for the task at hand, thereby enhancing its ability to make accurate predictions on previously unseen data**.**

7. Results and Discussions :

7.1. Comparison between different Classifiers :

To evaluate the performance of each classifier, we compare the results from four case studies. The table presents the outcomes (post-tuning), facilitating the identification of the most effective classifier through key metrics and rigorous statistical validation.

7.1.1. Evaluation de phase I :

Table 2: Phase I Evaluation

	LR	DT	SVM	RF	KNN	NB
Accuracy	**0.83**	**0.82**	**0.86**	**0.86**	**0.71**	**0.81**
Precision	**0.72**	**0.78**	**0.79**	**0.79**	**0.59**	**0.63**
Recall	**0.76**	**0.62**	**0.76**	**0.76**	**0.29**	**0.94**
F1-Score	**0.74**	**0.69**	**0.78**	**0.78**	**0.39**	**0.75**

7.1.2. Evaluation de phase II :

Table 3: Phase II Evaluation

	LR	DT	SVM	RF	NB
Accuracy	**0.84**	**0.796**	**0.85**	**0.84**	**0.83**
Precision	**0.84**	**0.75**	**0.85**	**0.85**	**0.87**
Recall	**0.62**	**0.53**	**0.65**	**0.76**	**0.71**
F1-Score	**0.71**	**0.62**	**0.73**	**0.73**	**0.73**

7.2. Findings :

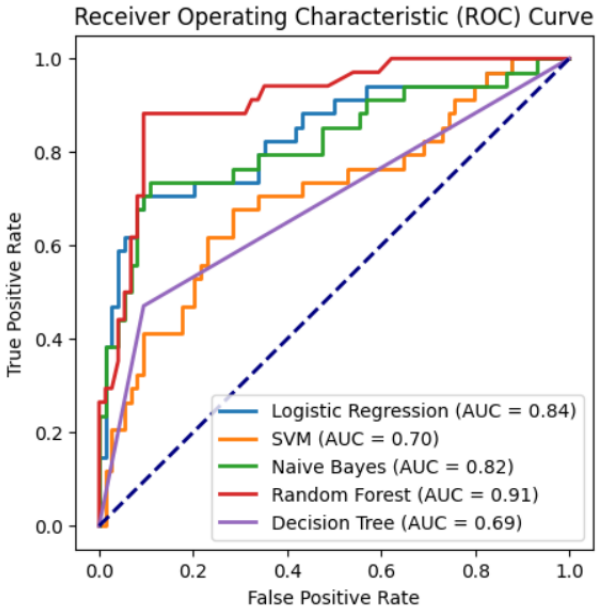

Figure 5: Roc Curve

The AUC is the area under the ROC curve, measuring the model's ability to distinguish between classes.

Interpretation of AUC:

AUC = 1.0: Perfect classifier.

AUC = 0.5: No discrimination (random chance).

Higher AUC values (closer to 1) indicate better performance.
Logistic Regression (AUC = 0.84): Performs well, showing a good balance between True Positive Rate and False Positive Rate.
SVM (AUC = 0.70): Moderate performance; not as effective as Logistic Regression or Random Forest.
Naive Bayes (AUC = 0.82): Performs better than SVM and slightly worse than Logistic Regression.
Random Forest (AUC = 0.91): The best-performing model in this comparison, with the highest AUC value.
Decision Tree (AUC = 0.69): The lowest-performing model among the five, with the least ability to discriminate between classes.

8. Discussion :

Random Forest is the most reliable model for your data based on the ROC curve, as it has the highest AUC.
Logistic Regression and Naive Bayes also perform well and are good alternatives if you prioritize simpler models.
SVM and Decision Tree show comparatively weaker performance, and they may require additional tuning or may not be suitable for this task.
Our comparative analysis of machine learning models for PCOS detection revealed varying performance levels across different metrics. Random Forest emerged as the leading classifier with an AUC of 0.91, demonstrating superior discrimination capabilities. Logistic Regression and Naive Bayes also showed robust performance (AUC = 0.84 and 0.82 respectively), making them viable alternatives when model simplicity is prioritized. While SVM and Decision Tree achieved moderate results (AUC = 0.70 and 0.69), their performance suggests potential limitations for this specific classification task.
The performance metrics table indicates consistent patterns across multiple evaluation criteria. Random Forest achieved balanced scores in accuracy (0.84), precision (0.85), and F1-score (0.73), while maintaining the highest recall (0.76) among all models. This balanced performance across metrics reinforces its position as the most reliable classifier for PCOS detection in our dataset.
These findings suggest that ensemble methods, particularly Random Forest, offer the most robust approach for PCOS classification, while simpler models like Logistic Regression remain competitive alternatives when interpretability is crucial.

9. Model Deployment :

The model has been successfully deployed, providing an innovative tool for the detection of Polycystic Ovary Syndrome (PCOS). Users are prompted to input a set of relevant symptoms into a structured form. Upon submission, the model processes these inputs using advanced algorithms to generate a personalized result. This result offers an insightful analysis based on the user's specific symptoms, helping to assess the likelihood of PCOS. By leveraging this model, users gain valuable preliminary insights into their health,

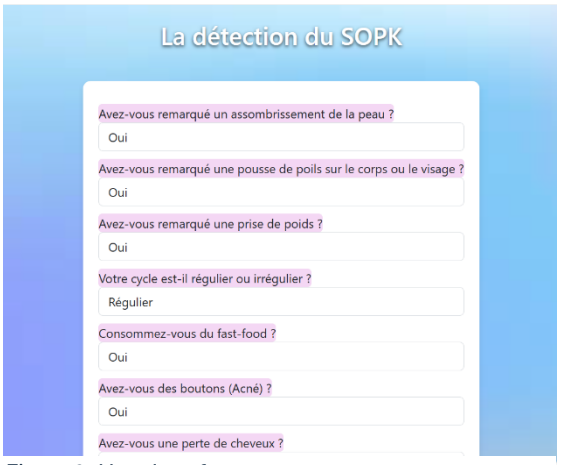

Figure 6: User Interface

facilitating informed decision-making for further medical consultation .

Figure 7: Result Interface

IV- Limitations and Role of the Model in Early PCOS Detection :

While the deployed model offers a valuable preliminary assessment tool for Polycystic Ovary Syndrome (PCOS), it is essential to acknowledge its limitations. The model relies on user-reported symptoms, which may not always encompass the full spectrum of clinical factors required for an accurate diagnosis. Ultrasound remains the most reliable and widely accepted method for diagnosing PCOS, as it allows for a more comprehensive evaluation of the ovaries and other relevant factors. Nevertheless, the purpose of this work is to provide an accessible first step for individuals, particularly women, to identify potential signs of PCOS early on. By facilitating the recognition of common symptoms, this model serves as a supportive tool, encouraging users to seek professional medical advice and timely diagnostic procedures.

V- References :

1.Escobar-Morreale H.F. Polycystic ovary syndrome: Definition, aetiology, diagnosis and treatment. Nat. Rev. Endocrinol. 2018;14:270–284. doi: 10.1038/nrendo.2018.24. [DOI] [PubMed] [Google Scholar

2.Norman R.J., Dewailly D., Legro R.S., Hickey T.E. Polycystic ovary syndrome. Lancet. 2007;370:685–697. doi: 10.1016/S0140-6736(07)61345-2. [DOI] [PubMed] [Google Scholar].

3.McCartney C.R., Marshall J.C. Polycystic ovary syndrome. N. Engl. J. Med. 2016;375:54–64. doi: 10.1056/NEJMcp1514916. [DOI] [PMC free article] [PubMed] [Google Scholar]