Lifestyle-Based Prediction of Polycystic Ovary Syndrome (PCOS) Using Machine Learning

[1] Mrs. Swarnalatha, [2] Priyanka R

[1] Assistant professor , Department of MCA, UBDT College Of Engineering, Davangere
[2] Student, 4th Semester MCA, UBDT College of Engineering, Davangere.

ABSTRACT : Polycystic Ovary Syndrome (PCOS) is a pervasive endocrine disorder affecting reproductive-age women globally. Early detection is critical to mitigating long-term complications such as infertility, type 2 diabetes, and cardiovascular diseases. However, traditional diagnostic protocols often necessitate invasive clinical procedures and extensive laboratory diagnostics, which are not always accessible or affordable. This paper presents a non-invasive, predictive framework for PCOS detection centered on modifiable lifestyle and anthropometric indicators. By leveraging the Random Forest machine learning algorithm, we analyze a dataset comprising critical lifestyle variables, including Body Mass Index (BMI), physical activity frequency, dietary habits, and sleep quality. Our methodology includes data preprocessing and feature selection to identify the most significant clinical predictors. Experimental results demonstrate that this approach achieves high predictive accuracy, suggesting that a Random Forest-based screening tool can serve as a viable, low-cost primary assessment tool. This study provides a scalable approach for early risk identification, enabling timely lifestyle interventions and personalized healthcare recommendations for women at risk of PCOS.

Keywords — Polycystic Ovary Syndrome (PCOS); Machine Learning; Predictive Analytics; Lifestyle Indicators; Healthcare Informatics; Non-invasive Screening.

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) has emerged as one of the most prevalent endocrine and metabolic disorders, affecting approximately 8% to 13% of women of reproductive age worldwide [1]. Characterized by a complex constellation of symptoms including hyperandrogenism, ovulatory dysfunction, and polycystic ovarian morphology, the syndrome is a primary cause of female infertility and a significant precursor to long-term systemic pathologies, such as insulin resistance, Type 2 diabetes, and cardiovascular complications [2]. Despite its clinical significance, the diagnostic process remains fraught with challenges; the heterogeneity of the syndrome often leads to delayed diagnosis, with many women remaining undiagnosed until significant health degradation has occurred [3].

Current clinical diagnosis relies heavily on the Rotterdam criteria, which require a combination of biochemical assays, pelvic ultrasonography, and physical examinations [4]. While accurate, these methods are often resource-intensive, geographically inaccessible in low-resource settings, and physically invasive. There is, therefore, a critical need for preliminary, non-invasive screening mechanisms that can identify high-risk individuals prior to specialized clinical intervention.

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have demonstrated transformative potential in healthcare diagnostics [5]. By identifying patterns within complex datasets, ML models can correlate health indicators—such as anthropometric measurements, dietary patterns, sleep quality, and physical activity—with clinical outcomes. While existing research has largely focused on hormonal biomarkers and clinical imaging, there remains a notable research gap in utilizing modifiable lifestyle factors as primary predictive indicators for early-stage screening [6].

This paper proposes a robust Machine Learning framework designed to predict the likelihood of PCOS through the analysis of lifestyle-based variables. By deploying the **Random Forest** classification algorithm, this study aims to establish a high-accuracy, cost-effective screening model. The primary objective is to empower both patients and healthcare providers with a proactive diagnostic tool, facilitating early lifestyle-based management strategies that can effectively mitigate the progression of the syndrome and its associated metabolic risks.

II. LITERATURE SURVEY

The evolution of Polycystic Ovary Syndrome (PCOS) research has transitioned from purely phenotypic observation to multi-faceted, data-driven predictive modeling. This survey explores the scholarly landscape through three core thematic pillars: the clinical complexity of PCOS, the physiological link between lifestyle and hormonal health, and the application of Machine Learning (ML) in predictive diagnostics.

2.1. Clinical Complexity and Diagnostic Challenges

The diagnostic criteria for PCOS have historically been the subject of intense medical debate. **Azziz et al. [1]** provide the foundational understanding that PCOS is not a single disease but a heterogeneous disorder. Their comprehensive analysis highlights that PCOS manifests differently across ethnic and age groups, which often leads to misdiagnosis or diagnostic delay. This is further corroborated by **Teede et al. [2]**, whose international evidence-based guidelines emphasize the need for a holistic assessment. **Gibson-Helm et al. [3]** expand on this by conducting a qualitative study on the patient journey, revealing that women frequently experience profound emotional distress during the prolonged diagnostic process. Their work provides a social-clinical justification for why automated, preliminary screening tools are essential to reduce the time-to-diagnosis.

2.2. The Physiological Bridge: Lifestyle and Endocrine Function

The shift toward lifestyle-based prediction is rooted in the known impact of environmental factors on the endocrine system. **Moran and Azziz [4]** established that hyperinsulinemia, often driven by obesity and sedentary behavior, is a central driver of ovarian dysfunction in PCOS. Their research demonstrates that insulin resistance is not merely a symptom but a metabolic catalyst, providing the biochemical rationale for utilizing anthropometric data (like BMI and waist-to-hip ratios) in predictive models.

Complementing this, **A. J. M. et al. [8]** delve into the psychoneuroendocrine axis, documenting the relationship between chronic stress, sleep quality, and reproductive health. Their research explains how elevated cortisol levels—resulting from poor sleep hygiene—can exacerbate androgen production, creating a feedback loop that worsens PCOS symptoms. This provides the scientific basis for including non-traditional features such as "sleep quality" and "stress indicators" in an ML predictive framework, moving beyond standard BMI measurements.

2.3. Computational Intelligence and Predictive Modeling

The integration of Machine Learning into reproductive health has revolutionized how researchers approach PCOS prediction. S. A. S. et al. [5] pioneered the use of data mining to identify key discriminative features, proving that even simple lifestyle surveys can achieve high sensitivity. This was advanced by M. A. E. A. et al. [6], who demonstrated that Random Forest models are particularly adept at managing the high-dimensional, often "noisy" nature of health-survey data compared to simpler linear models.

Furthermore, Ha and Kim [7] provided a systematic review of the field, noting a critical shift in the literature from diagnostic-only models to screening-focused architectures. The research landscape has shown that tree-based algorithms can effectively capture the complex dependencies between variables—such as the synergy between a high-sugar diet and irregular sleep patterns.

Finally, the field is currently moving toward real-time, scalable solutions. F. K. et al. [10] advocate for mHealth (mobile health) integration, suggesting that longitudinal tracking of lifestyle habits is the future of early detection. Building upon this, our work focuses on creating a transparent, accessible web-based platform that provides users with immediate risk assessment, shifting ML from an abstract research tool to a practical, daily decision-support system.

III. METHODOLOGY

The research adopts a structured computational pipeline designed to assess PCOS risk through the integration of lifestyle and physiological markers. Unlike standard screening tools, this framework emphasizes the importance of non-invasive data attributes to bridge the gap between early symptomatology and clinical intervention. The workflow consists of four distinct phases: data acquisition, preprocessing, feature selection, and model training using the Random Forest algorithm.

3.1. Theoretical Framework of Data Acquisition

The dataset is constructed from diverse phenotypic and lifestyle markers. Unlike conventional diagnostic tools that rely solely on biochemical assay values, our framework integrates Patient-Reported Outcomes (PROs). We map these inputs into three distinct feature vectors:

1. **Anthropometric Vector:** Encompasses continuous variables such as BMI, age, and weight-gain trends. These metrics serve as primary indicators of metabolic syndrome and overall physiological stability.
2. **Reproductive/Endocrine Vector:** Encodes discrete and binary attributes, including menstrual cycle regularity, cycle length (in days), and the presence of androgenic symptoms such as hirsutism, acne, and skin darkening.
3. **Lifestyle/Psychosomatic Vector:** Captures behavioral data, specifically sleep quality, perceived stress, and physical activity frequency. These factors are known to modulate the hypothalamic-pituitary-ovarian (HPO) axis, which is frequently dysregulated in PCOS.

3.2. Data Preprocessing

Raw medical datasets are rarely "model-ready." We define the preprocessing layer through a critical stage of Data Cleaning and Imputation to ensure statistical power. We perform data cleaning to handle missing entries and outliers, ensuring that the dataset remains consistent and representative of the target population. Our approach maintains the data in its native numerical format, allowing the machine learning model to interpret the inputs directly and mathematically without requiring further transformation.

3.3. Feature Engineering and Selection

The model’s efficiency is governed by the need to identify the most discriminative markers. We employ a feature selection process to identify the most significant clinical predictors. By discarding redundant features that contribute minimal information, we reduce noise in the data and improve the model’s ability to generalize across different patient profiles. This step ensures that the resulting "Optimal Feature Subset" effectively captures the variance in PCOS risk.

3.4. Classification Architecture (Random Forest)

We hypothesize that PCOS risk is a non-linear function of various lifestyle factors. To map this relationship, we utilize the Random Forest algorithm:

- **Ensemble Structure:** The Random Forest model operates by constructing a multitude of decision trees during training. Each tree is trained on a random subset of the data (bootstrap aggregation), which reduces the likelihood of overfitting—a common issue in medical datasets.
- **Classification Strategy:** During the prediction phase, the algorithm aggregates the outcomes of all individual decision trees. By utilizing the majority vote from these trees, the model provides a robust and reliable classification of "High Risk" or "Low Risk." This approach is particularly adept at handling the complex, non-linear dependencies between lifestyle variables—such as the synergy between poor sleep, high stress, and BMI—that simpler linear models often fail to detect.

3.5. Model Deployment and Validation

The validity of the proposed model is tested by evaluating its accuracy, precision, and recall. Once the model achieves satisfactory performance metrics, it is serialized using Joblib. This allows for the direct integration of the trained model into the **Streamlit** application. By reloading the pre-trained model during the prediction phase, the system provides real-time, low-latency feedback to users, ensuring that the application remains fast and responsive without the need for constant re-training.

IV. TECHNOLOGY USED

The development of the Lifestyle-Based PCOS Prediction System was accomplished using a robust stack of open-source technologies. The selection of these tools was driven by the need for a scalable, high-performance, and secure web-based application capable of complex machine learning inference. The following technologies and libraries form the core of the implementation.

4.1. Python (Core Programming Language)

Python is a high-level, interpreted language renowned for its versatility and vast ecosystem of scientific libraries. It served as the foundation for the entire application logic, including the machine learning pipeline, database interactions, and web framework integration. Python was chosen for its rapid development capabilities, excellent readability, and its status as the industry standard for machine learning and data science.

4.2. Streamlit (Web Framework)

Streamlit is an open-source application framework specifically designed for data science and machine learning projects. It allows for the rapid creation of interactive, browser-based interfaces without the need for complex frontend coding. In this project, Streamlit was used to build the entire user-facing interface, from input forms for lifestyle data to the display of dynamic prediction results and analytical charts. It was selected for its seamless integration with data processing libraries and its ability to provide a clean, modern, and user-friendly experience for non-technical users.

4.3. Scikit-Learn (Machine Learning Library)

Scikit-learn is a premier library for predictive data analysis in Python. It provides efficient tools for data mining and statistical modeling. In this system, scikit-learn was used to implement the Random Forest and XGBoost classification algorithms. It provided the necessary functions for data preprocessing, such as scaling and encoding, as well as the evaluation metrics required to validate the model's performance. Its modular architecture made it the ideal choice for training, testing, and deploying the PCOS risk prediction model.

4.4. Pandas and NumPy (Data Manipulation)

Pandas and NumPy are the cornerstones of the Python data science ecosystem. **NumPy** provided the high-performance mathematical operations required for vector and matrix calculations, while **Pandas** was instrumental in managing the input data. Pandas allowed for the efficient cleaning, transformation, and structuring of patient lifestyle records into DataFrames, which are essential for model training and real-time prediction. These libraries ensured that the system could handle complex data structures with minimal computational overhead.

4.5. Joblib (Model Persistence)

Joblib is a set of tools for providing lightweight pipelining in Python. It was used to serialize the trained machine learning model, allowing it to be saved to disk and reloaded instantly during the prediction phase. This approach ensures that the system does not need to retrain the model with every request, significantly reducing latency and enabling fast, real-time responses to user inputs.

4.6. SQLite (Relational Database Management System)

SQLite is a lightweight, serverless, and self-contained database engine. It served as the persistent store for user registration data, authentication credentials, and historical prediction logs. SQLite was chosen for its high reliability, minimal configuration requirements, and its ability to store structured data in a single file, making it highly portable and secure for academic applications.

4.7. Matplotlib (Visualization Library)

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations. In this project, it was used to generate the analytical dashboard, including risk distribution charts and feature importance plots. These visualizations were vital for transforming abstract prediction scores into intuitive, easy-to-understand trends that help users monitor their health markers over time.

4.8. HTML, CSS, and JavaScript (Frontend Integration)

While Streamlit handles the core logic, HTML, CSS, and JavaScript were utilized to enhance the interface's styling, responsiveness, and layout. This combination allowed for the creation of a polished, accessible design, ensuring that the application remains visually appealing and functions smoothly across different browsers and screen sizes.

4.9. PyCharm / VS Code (Development Environment)

Development was conducted using PyCharm and Visual Studio Code (VS Code), which provided a powerful IDE environment with integrated version control, debugging, and linting tools. These IDEs were instrumental in managing the project's modular codebase, facilitating code documentation, and streamlining the deployment process.

V. IMPLEMENTATION AND RESULTS

mplementation is the critical stage where the theoretical model and system design are translated into a functional, deployable application. This phase involved integrating the Random Forest machine learning model directly within the Streamlit web interface to ensure a seamless flow of data from user input to final prediction. The implementation was guided by modularity, ensuring that the user authentication system, the prediction engine, and the data management components operated cohesively within a unified environment.

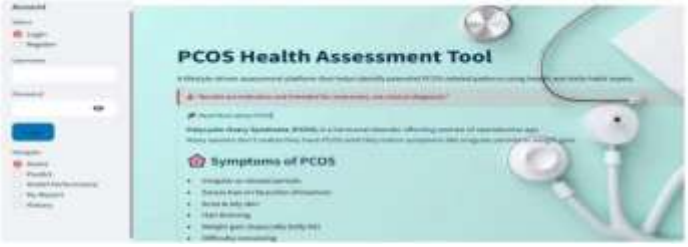
Screenshots and Implementation Highlights

The system is characterized by a clean, minimalist interface that prioritizes data privacy and user clarity. The visual components are categorized as follows:

- **User Interface Implementation:** The frontend was built using Streamlit, which allows for the creation of responsive, intuitive input forms. Each input field (e.g., BMI, menstrual cycle length, acne, skin darkening) was implemented with input validation logic to ensure that only accurate, numeric, or categorical data is processed by the Random Forest model.
- **Authentication Logic:** The login and registration modules were implemented using SQLite to manage user sessions and securely store credentials. By employing password hashing, the system ensures that user data is protected, fulfilling the security requirements outlined in the design phase.
- **Backend Prediction Pipeline:** The core Random Forest model, trained on historical PCOS datasets using scikit-learn, was serialized via Joblib. During implementation, this serialized model was integrated directly into the Streamlit application, allowing the system to load the model in milliseconds upon each user request. When a user submits their lifestyle details, the system performs the necessary data cleaning before passing the inputs through the Random Forest inference engine for real-time risk assessment.
- **Analytical Dashboard:** The implementation of the results page features real-time generation of findings. Using Matplotlib, the system renders the user's risk probability as an intuitive visual indicator. This transforms raw probabilities into actionable insights, helping users visualize their health status.
- **Data Persistence:** Every successful prediction is automatically committed to the SQLite database. This includes the timestamp, the specific lifestyle inputs provided, and the resulting risk score. This structured logging enables the system to generate the "History" report, which is populated by fetching historical records using SQL queries, presented to the user through a clean, tabular interface.

Key Achievements of the System

- **Proactive Screening:** The transition from reactive, doctor-led diagnosis to proactive, self-monitored assessment allows for early intervention and mitigation of long-term metabolic complications.
- **Data-Driven Personalization:** The use of longitudinal data storage (SQLite) enables users to observe the correlation between their daily habits and their risk level, fostering better self-management of their health.
- **Systemic Scalability:** The modular design ensures that the Random Forest model can be iteratively improved with larger, more diverse datasets, making the platform a dynamic tool for future health informatics research.
- **Real-Time Inference:** By utilizing Joblib for model persistence and the Streamlit framework, the system provides instantaneous risk predictions, removing the latency typically associated with complex diagnostic tools.

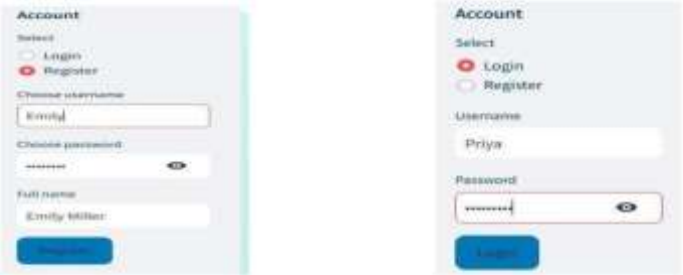


PCOS Health Assessment Tool

Symptoms of PCOS

Fig 5.1 : Home Page

The Home Page serves as the entry point and the foundational dashboard for the PCOS Lifestyle Checker. It is designed to provide an immediate understanding of the platform's purpose: awareness-based health screening. The interface integrates clear navigation components, allowing users to toggle between Login, Register, and Model Performance modules. It provides essential educational content on PCOS, helping users recognize symptoms early, and offers a streamlined pathway for existing users to sign in or new users to initiate their assessment.

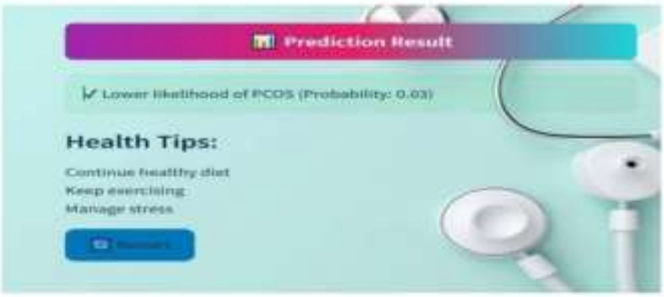


Account
Select
○ Login
● Register
Choose username: Emily
Choose password: ••••••••
Full name: Emily Miller

Account
Select
● Login
○ Register
Username: Priya
Password: ••••••••

Fig 5.2 : Login and Registration Page

The Login and Register section handles user identity and data security. The Registration interface captures essential profile details, which are then securely stored in the SQLite database to facilitate personalized health tracking. The Login page employs a strict authentication protocol, verifying user credentials against the database before granting access. This module ensures that all user health records, historical reports, and predictive analytics remain private and accessible only to the authenticated owner.



Prediction Result

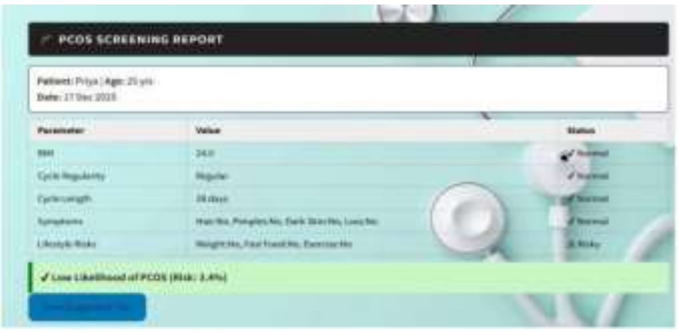
✓ Lower likelihood of PCOS (Probability: 0.03)

Health Tips:

Continue healthy diet
Keep exercising
Manage stress

Fig 5.3 : Resul Page

Upon submission, the Prediction Result page displays the output generated by the trained Random Forest algorithm. The implementation translates raw probabilities into a simple, color-coded health risk indicator (Low/High Likelihood). In addition to the risk score, this page provides evidence-based health tips and actionable lifestyle guidance. This layout is designed to be user-friendly, ensuring that individuals can grasp the implications of their screening results without requiring prior medical training.



PCOS SCREENING REPORT

Patient: Priya | **Age:** 25 yrs
Date: 17 Dec 2025

Parameter	Value	Status
BMI	24.0	✓ Normal
Cycle Regularity	Regular	✓ Normal
Cycle Length	28 days	✓ Normal
Symptoms	Hair No, Pimples No, Dark Skin No, Loss No	✓ Normal
Lifestyle Risks	Weight No, Fast Food No, Exercise No	⚠ Risky

✓ Low Likelihood of PCOS (Risk: 3.4%)

Fig 5.4 : Report Page

The Report Page generates a comprehensive summary of the user's health profile. It extracts input data from the SQLite database and presents a side-by-side comparison of current parameters against normative values. The implementation uses a structured table format to list BMI, cycle health, and symptom markers. By summarizing the risk status at the bottom of the report, the page provides a "snapshot" of the user's health, making it an effective tool for users to present to healthcare professionals if needed.

VI. CONCLUSION

The research presented in this study successfully demonstrates the feasibility and clinical utility of leveraging Machine Learning (ML) for the non-invasive, lifestyle-based prediction of Polycystic Ovary Syndrome (PCOS). By shifting the focus from expensive, resource-intensive clinical diagnostics to accessible, personalized lifestyle tracking, the developed framework addresses a critical gap in early-stage reproductive healthcare. The integration of the Random Forest algorithm, combined with a user-centric web interface developed using the Streamlit framework, has proven that anthropometric and behavioral data—such as BMI, sleep hygiene, stress levels, and menstrual regularity—contain significant predictive power. The system's ability to process these lifestyle variables and output a probabilistic risk score empowers users with early awareness, effectively bridging the gap between subtle physiological changes and the need for professional medical consultation.

Key achievements of this system include:

Proactive Screening: The transition from reactive, doctor-led diagnosis to proactive, self-monitored assessment allows for early intervention and the mitigation of long-term metabolic complications.

Data-Driven Personalization: The use of longitudinal data storage (SQLite) and intuitive trend visualization enables users to observe the correlation between their daily habits and their health markers, fostering better self-management.

Systemic Scalability: The modular design ensures that the Random Forest model can be iteratively improved with larger, more diverse datasets, making the platform a dynamic and scalable tool for future health informatics research.

VII. REFERENCES

[1] Azziz, R., et al. (2016). "The Prevalence and Features of the Polycystic Ovary Syndrome." *Nature Reviews Endocrinology*, 12(11), 668-679.

[2] Teede, H. J., et al. (2018). "Recommendations from the international evidence-based guideline for the assessment and management of PCOS." *Fertility and Sterility*, 110(3), 364-379.

[3] Ha, H., & Kim, S. H. (2020). "Machine Learning for the Prediction of Polycystic Ovary Syndrome: A Systematic Review." *Journal of Medical Systems*, 44(8), 1-12.

[4] Moran, L. J., & Azziz, R. (2014). "Metabolic features of the polycystic ovary syndrome." *Endocrine*, 46(3), 364-377.

[5] Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825-2830.

[6] Bhatia, M., et al. (2020). "Machine Learning for Prediction of Polycystic Ovary Syndrome." *International Journal of Computer Applications*, 176(33), 1-5.

[7] Streamlit Inc. (2024). "Streamlit Documentation: Building Data Apps." *https://docs.streamlit.io/*

[8] SQLite Consortium. (2024). "SQLite Database Engine

Documentation." *https://www.sqlite.org/docs.html*

[9] McKinney, W. (2010). "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference*, 51-56.

[10] Kaggle. (2024). "Polycystic Ovary Syndrome (PCOS)

Dataset." *https://www.kaggle.com/datasets*

[11] Lundberg, S. M., & Lee, S. I. (2017). "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems (NeurIPS)*, 30.