A Preventive Machine Learning Framework for Multi-Level PCOS Risk Evaluation and Lifestyle-Centered Wellness Support

Priyadharshni R[1], Keerthana G[2], Rajeshwari T[3]
[1,2] *M.Sc. in Computer Science, Sathyabama Institute of Science and Technology, Chennai*
[3] *Assistant Professor, Sathyabama Institute of Science and Technology, Chennai*
doi.org/10.64643/IJIRTV12I10-194344-459

***Abstract*—Polycystic Ovary Syndrome (PCOS) is a widespread endocrine condition affecting women during their reproductive years. The disorder is influenced not only by biological mechanisms but also by lifestyle behaviors and psychological stress. In many cases, the condition remains unnoticed during its early stages, which may lead to more severe health complications over time. The present study proposes a preventive digital framework that employs machine learning techniques to estimate individual PCOS risk levels and provide lifestyle-based wellness support. The framework integrates multiple categories of user-reported information, including menstrual cycle characteristics, physical symptoms, lifestyle behaviors, and perceived stress indicators. Based on these inputs, the system classifies individuals into different levels of potential risk such as early, moderate, and high risk. In addition, the study examines the relationship between stress patterns and hormonal imbalance that may contribute to PCOS progression. Unlike conventional systems that concentrate mainly on clinical diagnosis after symptoms become evident, the proposed approach emphasizes preventive awareness and behavioral guidance. The framework is designed as a supportive decision-assistance tool rather than a replacement for professional medical diagnosis. Experimental analysis suggests that machine learning-assisted wellness platforms can contribute to early awareness and encourage healthier lifestyle practices for individuals who may be vulnerable to PCOS.**

***Index Terms*—PCOS, Machine Learning Models, Risk Evaluation, Preventive Wellness, Lifestyle-Centered Healthcare.**

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is considered one of the most prevalent hormonal disorders affecting women of reproductive age. The condition is characterized by irregular menstrual cycles, metabolic imbalance, and disturbances in reproductive hormones. Individuals affected by PCOS frequently experience symptoms such as irregular menstruation, excessive weight gain, acne, hair loss, insulin resistance, and chronic fatigue. In addition to these physical manifestations, many individuals also report emotional stress and psychological instability, suggesting that the condition is influenced by both physiological and behavioral factors.The number of PCOS cases has increased in recent decades, which may be linked to changing lifestyle patterns such as reduced physical activity, poor dietary habits, insufficient sleep, and increased psychological stress. Stress is particularly important because it may interfere with hormonal regulation through its influence on the hypothalamic–pituitary–ovarian axis. Despite this connection, most clinical monitoring systems still emphasize laboratory tests and imaging techniques rather than behavioral or psychological indicators.

Early recognition of potential risk factors is essential for preventing the long-term consequences associated with PCOS. If the condition remains unmanaged, it may lead to complications such as metabolic syndrome, type-2 diabetes, cardiovascular disorders, and infertility-related problems. Although modern medical technologies allow accurate diagnosis, many healthcare systems identify the condition only after symptoms become clearly visible.A major limitation of many existing predictive models is that they primarily focus on binary classification—determining whether a patient has PCOS or not—using clinical datasets. Less attention has been given to stage-wise risk assessment, stress-related influences, and lifestyle-based preventive strategies. As a result, the potential role of digital tools in promoting early awareness and lifestyle modification remains underutilized.

To address this gap, the present research introduces a machine learning-based wellness support system that evaluates PCOS risk patterns using a combination of physiological indicators and lifestyle factors. The system analyzes menstrual history, observable physical symptoms, behavioral habits such as sleep and diet, and self-reported stress indicators. Logistic regression is selected as the primary predictive technique because of its interpretability and suitability for healthcare-related decision-support systems.It is important to emphasize that the framework is not intended to function as a medical diagnostic system. Instead, it aims to increase awareness, guide individuals toward healthier lifestyle choices, and encourage consultation with healthcare professionals when necessary. By integrating predictive analytics with lifestyle-centered recommendations, the proposed system contributes to proactive health monitoring and improved reproductive health awareness.

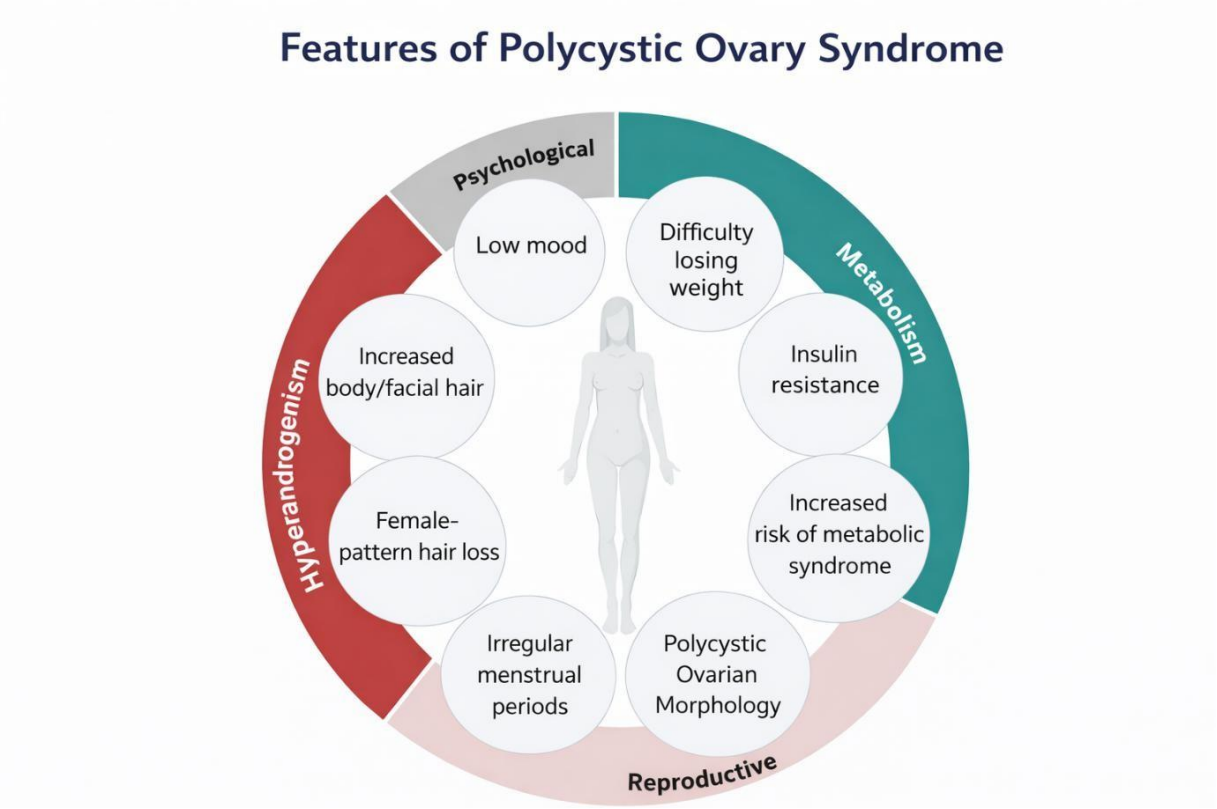

Fig. 1. Illustration of common manifestations associated with polycystic ovarian syndrome (PCOS), including reproductive irregularities, metabolic symptoms, hyperandrogenic features, and psychological stress factors.

II. LITERATURE REVIEW

R. Jayanthi, S. S. Nithya, and S. K. Sreeja (2019) examined the application of machine learning algorithms to detect PCOS using clinical health parameters such as menstrual irregularities, body mass index, and hormonal indicators. Their findings suggested that automated predictive systems can support early identification of PCOS cases. However, the study primarily relied on medical datasets and did not consider lifestyle behaviors or psychological stress as contributing factors.

S. J. Anitha, P. S. R. Reddy, and K. R. Rao (2020) explored the use of machine learning methods for predicting PCOS from patient medical records. Their analysis incorporated variables such as age, body mass index, menstrual irregularities, and hormonal indicators. Although the models demonstrated encouraging prediction accuracy, the research focused mainly on diagnostic classification rather than preventive health monitoring.

A. Denny, S. Rani, and R. Karthik (2021) proposed a predictive model for early PCOS identification using selected metabolic and reproductive health indicators. Their work confirmed that machine learning algorithms can assist in identifying potential PCOS cases. Nevertheless, the study emphasized

clinical and laboratory data, while behavioral and stress-related aspects were not integrated into the model.

Md. Mahbubur Rahman, Ashikul Islam, Forhadul Islam, Mashruba Zaman, Md. Rafiul Islam, Md. Shahriar Alam Sakib, and Hafiz Md. Hasan Babu (2023) proposed web-based systems that apply machine learning algorithms for PCOS prediction through interactive platforms. These systems evaluated various algorithms and identified models such as Random Forest and AdaBoost as effective predictors. While these platforms improved accessibility and user interaction, they still concentrated mainly on diagnostic prediction rather than preventive lifestyle guidance.

Aunik Hasan Mridul, Nowreen Ahsan, Syeda Sadia Alam, Sonia Afrose, Zakia Sultana, and Md. Tanvir Mahmud Kafi (2024) examined both traditional machine learning algorithms and deep learning models for PCOS prediction. Ensemble models achieved high levels of predictive accuracy when trained on clinical datasets. Despite their promising performance, these models largely focused on disease classification rather than incorporating behavioral and environmental factors.

Mehtap Agirsoy and Matthew A. Oehlschlaeger (2025 demonstrated that machine learning models such as XGBoost and neural networks can effectively detect PCOS. However, these studies largely emphasize diagnostic accuracy instead of exploring preventive risk evaluation or lifestyle-based health monitoring.

Overall, existing research demonstrates the potential of machine learning for PCOS prediction, but most studies prioritize medical diagnosis using clinical datasets. Few systems incorporate lifestyle behaviour, psychological stress, and preventive wellness strategies into predictive frameworks. This gap highlights the importance of developing integrated models that combine medical indicators with lifestyle and stress-related parameters.

II. EXISTING SYSTEM

Currently, the diagnosis of Polycystic Ovary Syndrome is primarily performed through clinical examination, ultrasound imaging, and laboratory hormone analysis. Medical professionals typically follow established diagnostic criteria that evaluate ovulation patterns, androgen levels, and ovarian morphology. While these procedures provide reliable confirmation of the condition, they are generally applied only after noticeable symptoms develop. Digital health tools have been introduced to assist individuals in monitoring reproductive health. Many mobile applications allow users to record menstrual cycles, weight fluctuations, and basic wellness indicators. However, most of these platforms function only as tracking systems rather than predictive analytical tools.

Although some research studies have applied machine learning techniques for PCOS prediction, these models are often developed in controlled research environments and are rarely integrated into practical user-oriented applications. In many cases, the systems lack interactive dashboards, real-time data analysis, and personalized health guidance. Another limitation of existing approaches is the minimal attention given to lifestyle and psychological factors that may contribute to hormonal imbalance. Elements such as sleep quality, physical inactivity, unhealthy dietary habits, and prolonged psychological stress are widely recognized as influencing metabolic disorders, yet they are rarely incorporated into digital predictive systems.

Therefore, current systems either rely on traditional clinical diagnosis without predictive digital support or provide basic symptom tracking without intelligent risk analysis. This creates a need for a more comprehensive system that combines lifestyle monitoring, stress assessment, and machine learning-based prediction in a user-friendly platform.

III. PROPOSED SYSTEM

The proposed system introduces an intelligent web-based platform designed to evaluate potential PCOS risk through machine learning techniques. The primary objective of the platform is to allow individuals to enter relevant health information and receive an analytical assessment of possible risk levels along with lifestyle-based wellness guidance.The system begins with a secure registration process where users create an account and access the platform through authenticated login credentials. After logging in, users complete a structured questionnaire designed to gather information related to menstrual history, age, body

mass index, perceived stress level, sleep patterns, fatigue, dietary habits, and physical activity.

Once the information is submitted, the data are securely stored within a relational database system. These inputs are then processed by a machine learning model that has been previously trained on a labeled dataset related to PCOS indicators. The predictive model applies a multiclass logistic regression algorithm to estimate the likelihood of different PCOS severity stages.Based on the analysis, the system generates an output indicating a potential stage category such as early, moderate, or severe risk. The result is displayed immediately to the user and simultaneously stored in the database to support future monitoring.

Following the prediction process, the user is directed to a personalized dashboard that provides health guidance and wellness suggestions. The dashboard includes information on dietary habits, exercise recommendations, and stress-management strategies that may help individuals maintain hormonal balance and improve overall health. Returning users can directly access their dashboard and review previous predictions without repeating the entire data entry process. This feature supports continuous monitoring and allows individuals to track potential changes in their health patterns over time.Through the integration of web technologies, database management systems, and machine learning algorithms, the proposed framework aims to provide an accessible platform for early awareness and preventive health monitoring related to PCOS.

IV. METHODOLOGY

This study focuses on developing a machine learning-based prediction model to identify the severity stage of Polycystic Ovary Syndrome (PCOS) using structured health and lifestyle data. The main objective of the model is to classify individuals into different PCOS risk stages by analyzing both physiological indicators and stress-related lifestyle factors. The approach emphasizes how daily habits and psychological stress contribute to hormonal imbalance, which plays a crucial role in PCOS progression.

A. Dataset Description

For model development, a structured dataset containing approximately 3500 records was used. Each record represents an individual profile consisting of multiple attributes associated with reproductive health and metabolic condition. The dataset includes features such as age, menstrual cycle length, skipped periods, body mass index (BMI), acne, hair fall, weight gain, stress level, sleep hours, fatigue, long sitting duration, irregular meals, junk food consumption, and physical activity levels. The target variable represents categorized PCOS risk stages. These stages were encoded numerically to support multiclass classification. Before training the model, the dataset was examined for consistency and basic preprocessing steps were performed to ensure that the data was structured properly for machine learning implementation. The dataset used in this project was organized specifically for academic research and experimental validation of PCOS stage prediction.

B. Analysis of Stress-Related Features

PCOS is widely associated with endocrine and metabolic imbalance, and stress is considered one of the influencing factors in hormonal disruption. Therefore, special attention was given to stress-related variables such as stress level, sleep duration, fatigue, physical inactivity, irregular food habits, and junk food intake.

To understand the relationship between these parameters and PCOS risk, correlation analysis was conducted. The correlation coefficient was calculated to measure the strength and direction of association between each stress-related feature and the target variable. The Pearson correlation coefficient is mathematically expressed as:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Where Xrepresents a selected feature (for example, stress level) and Yrepresents the PCOS risk stage.

Features showing meaningful correlation were retained in the final feature set. Positive correlation values indicated that higher stress or unhealthy lifestyle patterns were associated with increased PCOS severity, whereas protective behaviors such as sufficient sleep and physical activity showed relatively lower risk association.

C. Model Development Using Logistic Regression

After feature selection, the dataset was divided into training and testing subsets using an 80:20 ratio. The training set was used to allow the model to learn patterns between input variables and PCOS risk stages, while the testing set was used to evaluate

predictive performance on unseen data.

A multiclass Logistic Regression algorithm was chosen for classification because of its interpretability and suitability for medical decision-support systems. Logistic Regression does not directly predict class labels; instead, it estimates the probability of each class using a logistic (sigmoid) function.

The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Where:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

In this equation, $x_1, x_2, \ldots, x_n$represent the selected input features such as BMI, stress level, sleep hours, and other health indicators. The coefficients $\beta_1, \beta_2, \ldots, \beta_n$are learned during the training process and determine the weight or influence of each feature on the final prediction.

Each feature contributes proportionally to the calculated score z. The sigmoid function then transforms this score into a probability value between 0 and 1. For multiclass classification, the model assigns the PCOS stage corresponding to the highest probability.

D. Role of Stress in Final Prediction

In this model, stress-related variables contribute directly to the weighted linear combination. For example, higher stress levels, prolonged inactivity, and insufficient sleep increase the weighted score, thereby increasing the probability of predicting a higher severity stage. This aligns with biological understanding, as chronic stress may elevate cortisol levels, disturb insulin regulation, and influence androgen production, all of which are linked to PCOS symptoms. Thus, the model does not measure hormone levels directly but infers hormonal imbalance patterns indirectly through measurable behavioral and lifestyle parameters.

E. Model Evaluation and Output Generation

The trained model was evaluated using accuracy and classification metrics to ensure reliability. Accuracy is calculated as:

$$Accuracy = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

After successful evaluation, the trained model was serialized and saved for deployment. During runtime, when new user input is provided, the model receives the feature vector, computes probability scores for each PCOS stage, and returns the class with the highest probability as the final output. The output is represented as a numerical label corresponding to predefined categories such as early, moderate, or severe stage.Therefore, the final prediction result is derived from a probability-based classification process where physiological indicators and stress-related lifestyle parameters collectively influence the outcome. This structured methodology ensures that the prediction model reflects both medical reasoning and data-driven learning.

V. CONCLUSION

The present research introduced a web-based wellness support system designed to increase awareness about Polycystic Ovary Syndrome through machine learning-based risk assessment. The developed platform allows users to provide information regarding menstrual patterns, physical symptoms, lifestyle behaviors, and stress levels. These inputs are analyzed by a predictive model that estimates the potential severity stage of PCOS.A multiclass logistic regression model was implemented to perform the prediction task. Experimental evaluation demonstrated that the model is capable of identifying patterns associated with different PCOS risk levels with satisfactory performance. The system then communicates the predicted stage to users through an interactive dashboard and provides general lifestyle recommendations aimed at promoting healthier habits.

It is important to clarify that the proposed framework is intended as an awareness and wellness- support tool rather than a clinical diagnostic system. The results generated by the model should be considered informational and should not replace professional medical evaluation. Individuals experiencing symptoms related to PCOS should seek appropriate consultation with healthcare professionals.Overall, the study demonstrates that integrating machine learning algorithms with digital health platforms can provide valuable support for preventive healthcare. By encouraging early awareness and lifestyle improvement, such systems may contribute to better management of reproductive health conditions.

VI. FUTURE WORK

Although the developed system demonstrates promising functionality, several opportunities exist for further enhancement. Future studies may focus on integrating additional machine learning algorithms such as Random Forest, Support Vector Machines, and deep learning models in order to compare predictive performance and potentially improve classification accuracy. Training the model on larger and more diverse datasets would also enhance generalizability and reduce potential bias. Another important improvement involves incorporating clinical data sources, including hormonal test results, ultrasound findings, and metabolic indicators. The integration of these medical parameters could strengthen the predictive capability of the model and provide a more comprehensive evaluation of PCOS risk.

Longitudinal monitoring is another potential extension of the system. Allowing users to update their health information periodically would enable the system to track changes in risk patterns over time and assess the impact of lifestyle modifications.The chatbot feature included in the platform could also be enhanced using advanced natural language processing techniques**,** enabling more personalized and context-aware health guidance.

Additionally, converting the system into a mobile application would improve accessibility and encourage regular user engagement. Cloud-based deployment could further support scalable data processing and secure data storage.Finally, collaboration with healthcare institutions for clinical validation would strengthen the reliability and practical value of the system. Real-world testing could provide insights into how such digital tools may assist in preventive health management and early awareness of PCOS.

REFERENCES

[1] R. Jayanthi, S. S. Nithya, and S. K. Sreeja, "Machine learning approach for prediction of polycystic ovary syndrome," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 5, pp. 1234–1239, 2019.

[2] S. J. Anitha, P. S. R. Reddy, and K. R. Rao, "Prediction of polycystic ovary syndrome using machine learning techniques," *International Journal of Scientific Research in Computer Science*, vol. 5, no. 3, pp. 45–50, 2020.

[3] Denny, S. Rani, and R. Karthik, "A machine learning-based approach for early detection of polycystic ovary syndrome," *Journal of Medical Systems*, vol. 45, no. 7,pp. 1–10, 2021.

[4] Md. M. Rahman, A. Islam, F. Islam, M. Zaman, M. R. Islam, M. S. A. Sakib, and H. M. H. Babu, "Empowering early detection: A web-based machine learning approach for PCOS prediction," *IEEE Access*, vol. 11, pp. 45678–45689, 2023.

[5] A.H. Mridul, N. Ahsan, S. S. Alam, S. Afrose, Z. Sultana, and M. T. M. Kafi, "Polycystic ovary syndrome disease prediction using traditional machine learning and deep learning algorithms," *Computers in Biology and Medicine*, vol. 167, pp. 107– 118, 2024.

[6] M. Agirsoy and M. A. Oehlschlaeger, "Machine learning-based diagnosis of polycystic ovary syndrome using clinical, biochemical, and ultrasound features," *Artificial Intelligence in Medicine*, vol. 149, pp. 102–112, 2025.

[7] R. Azziz, E. Carmina, D. Dewailly, E. Diamanti-Kandarakis, H. F. Escobar-Morreale, and W. Futterweit, "The androgen excess and PCOS society criteria for the polycystic ovary syndrome: The complete task force report," *Fertility and Sterility*, vol. 91, no. 2, pp. 456–488, 2009.

[8] J. A. March, D. R. Moore, K. E. Willson, A. M. Phillips, J. A. Lippman, and D. E. Peterson, "The impact of lifestyle modification on polycystic ovary syndrome: A systematic review," *Human Reproduction Update*, vol. 20, no. 3, pp. 347–363, 2014.

[9] M. R. Teede, A. Misso, M. L. Costello, B. Dokras, J. Laven, and L. Moran, "Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome," *Clinical Endocrinology*, vol. 89, no. 3, pp. 251–268, 2018.

[10] S. M. Johnson, K. L. Porter, and R. P. Smith, "Machine learning applications in women's health: A review," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 95– 110, 2021.

[11] E. Diamanti-Kandarakis, R. J. Dunaif, and R. J. Nestler, "Pathophysiology and types of polycystic ovary syndrome," *Endocrine Reviews*, vol. 38, no. 3, pp. 321–350, 2017.

[12] L. Moran, H. Teede, and J. Vincent, "Lifestyle changes in women with polycystic ovary

syndrome," *Cochrane Database of Systematic Reviews*, no. 7, Article ID CD007506, 2011.

[13] S. K. Patel, R. K. Shah, and M. A. Joshi, "Stress and its influence on endocrine disorders: A clinical perspective," *Journal of Endocrinological Investigation*, vol. 42, no. 9, pp. 1031–1040, 2019.

[14] T. J. Deo, M. S. Kumar, and P. R. Rao, "Artificial intelligence and machine learning in preventive healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1624–1635, 2020.

[15] K. H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 719–731, 2018.