

A Data-Driven Predictive Model for PCOS Risk Assessment and Personalized Lifestyle Intervention

Abia P¹

Department of CSE(AIML),
East Point College of Engineering
and technology, Bangalore, India
abiapaulraj@gmail.com

Arpitha²

Department of CSE(AIML),
East Point College of Engineering
and technology, Bangalore, India
lep23ci005@eastpoint.ac.in

Meghana K B³

Department of CSE(AIML),
East Point College of Engineering
and technology, Bangalore, India
lep23ci029@eastpoint.ac.in

P. Divya⁴

Department of CSE(AIML),
East Point College of Engineering
and technology, Bangalore, India
lep23ci036@eastpoint.ac.in

Priyanka.R⁵

Department of CSE(AIML),
East Point College of Engineering
and technology, Bangalore, India
lep23ci041@eastpoint.ac.in

Shilpa S P⁶

Department of CSE(AIML),
East Point College of Engineering
and technology, Bangalore, India
lep23ci055@eastpoint.ac.in

Abstract—Polycystic Ovary Syndrome (PCOS) is a common hormonal disorder affecting women of reproductive age and is associated with infertility, metabolic problems and psychological stress. Many existing studies focus on its causes, lifestyle factors, mental health effects and prediction using machine learning models such as neural networks, Convolutional neural network (CNN) and hybrid algorithms. The proposed system combines medical data, lifestyle factors and psychological indicators to improve PCOS prediction. It integrates hybrid deep learning models with explainable AI to provide accurate, transparent and early diagnosis. Additional features such as stress levels, sleep patterns, diet are included to enhance prediction and support personalized healthcare recommendations. This system also provides risk assessment, early warning alerts and personalized lifestyle suggestions to help users manage symptoms and reduce long term complications. By integrating multiple data sources and advanced analytics, the model aims to assist doctors in decision making and promote early detection and better management of PCOS.

Keywords—PCOS, Hybrid deep learning, Machine learning, Explainable AI (XAI), Early diagnosis, Risk Prediction

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a common hormonal disorder that affects women of reproductive age and is associated with irregular menstrual cycles, weight gain, acne, infertility, metabolic complications, and psychological problems. PCOS can also impact reproductive and mental health, with studies highlighting the relationship between PCOS, stress, anxiety, and depression [3],[4],[6],[7]. Despite how widespread it is, early detection remains difficult due to variations in symptoms, limited awareness, and restricted access to specialized healthcare. Delayed diagnosis can increase the risk of long-term complications, which underscores the need for early prediction and preventive healthcare [3][8]. Traditional PCOS diagnosis relies on clinical symptoms, medical history, hormonal tests, and ultrasound examinations. Machine learning techniques have been increasingly explored to identify patterns in medical data and support PCOS prediction. Previous studies have looked into machine learning-based PCOS detection systems [1], neural network prediction models [2], and deep learning approaches for automated diagnosis [9]. Supervised learning algorithms, particularly Support Vector Machines (SVM), can analyze structured clinical data, while Convolutional Neural Networks (CNN) can extract relevant features from ultrasound images. These techniques provide a foundation for building automated PCOS prediction systems. The proposed system integrates clinical information, hormonal parameters, ultrasound images, and lifestyle factors to deliver a comprehensive PCOS risk assessment. Clinical data such as age, weight, menstrual history, symptoms, and medical

history are processed using the Support vector machine (SVM) model, while optional ultrasound images are analyzed using the CNN model. Hormonal parameters, including Luteinizing hormone (LH), Follicle stimulating hormone (FSH), insulin, and testosterone, provide additional information when available. A prediction fusion module combines the outputs of both models to generate a final classification of Detected or Not Detected, along with a confidence score. Prior studies on PCOS-related lifestyle factors and prediction approaches highlight the value of combining medical and lifestyle information [1],[5]. The system also incorporates Explainable Artificial Intelligence (XAI) to improve the transparency of predictions and identify key contributing factors. Based on the prediction result, symptoms, and lifestyle information, users are classified into Low, Moderate, or High Risk categories. The system further provides personalized recommendations related to diet, exercise, sleep, and stress management. Research examining stress, anxiety, and depression among women with PCOS supports the importance of addressing psychological well-being alongside physical health [4],[6]. These features are intended to support preventive healthcare and help users manage their symptoms more effectively. The primary goal of this research is to develop a user-friendly, data-driven framework for early PCOS risk assessment and personalized lifestyle intervention. The system includes report generation and secure data storage to retain prediction results, risk levels, and recommendations for future reference. By combining clinical data analysis, machine learning, optional ultrasound image processing, and personalized healthcare suggestions, the proposed framework aims to assist healthcare professionals in decision-making and encourage early detection and better PCOS management. The research also supports Sustainable Development Goal 3 (Good Health and Well-being) by promoting preventive healthcare and women's health [3],[8].

II. RELATED WORK

Recent studies have explored machine learning and deep learning approaches for PCOS prediction and diagnosis. Machine learning systems using algorithms such as Support Vector Machine (SVM) and Random Forest have been developed for symptom-based PCOS detection, demonstrating the potential of data-driven methods for early identification [1]. Neural network-based approaches, including Back Propagation Neural Networks (BNN), have also been investigated for PCOS prediction using clinical data, though their performance depends on sufficient training data and model interpretability remains a challenge [2]. Clinical and lifestyle studies have identified hormonal

imbalance, insulin resistance, obesity, diet, and physical inactivity as key factors associated with PCOS [3], [5].

Psychological factors have also been examined, with studies reporting associations between PCOS and higher levels of stress, anxiety, and depression [4], [6]. Population-focused research has further looked at the prevalence and effects of PCOS among women of reproductive age, particularly its links to infertility and metabolic complications [7], [8]. More recently, explainable CNN and hybrid deep learning models have been explored for PCOS prediction using medical images, with XAI techniques such as SHAP and LIME offering insight into the factors that influence model predictions [9]. However, existing approaches generally focus on individual aspects such as clinical prediction, image analysis, lifestyle factors, or psychological characteristics. These limitations motivate the development of an integrated framework that combines clinical and optional ultrasound data for PCOS prediction, followed by risk assessment and personalized lifestyle recommendations.

III. WORKING METHODOLOGY

The proposed methodology for PCOS risk assessment and personalized lifestyle intervention follows a systematic pipeline that covers data collection, preprocessing, machine learning and deep learning-based prediction, prediction fusion, risk assessment, personalized recommendation, and report generation. The system takes clinical, lifestyle, and hormonal information as its primary input, with ultrasound images available as an optional addition. Clinical data are processed using a Support Vector Machine (SVM), while ultrasound images are analyzed through a Convolutional Neural Network (CNN). When both data sources are provided, their prediction outputs are combined via a hybrid fusion module to produce the final PCOS prediction and confidence score. This prediction is then used to classify risk levels and generate personalized lifestyle recommendations.

A. Data Collection

The data collection module serves as the initial stage of the proposed system. The user provides relevant personal, clinical, hormonal, and lifestyle information through a mandatory clinical questionnaire. As illustrated in Fig.1, collected parameters include age, height, weight, menstrual-cycle information, symptoms, medical history, lifestyle habits, and stress level. Additional hormonal information such as LH, FSH, insulin, and testosterone can be provided when available. An ultrasound image can also be uploaded as an optional input to provide additional information for prediction.

The system is designed to continue prediction using the available clinical information when optional ultrasound or hormone data are not provided. Thus, the proposed framework can utilize different combinations of available health information.

B. Preprocessing of Data

The collected data are divided into tabular clinical data and ultrasound image data, and each type is processed separately before prediction. The preprocessing stage improves the quality and consistency of the input data and prepares them for the respective prediction models.

For tabular data, missing values are handled, unusual or outlier values are identified, and categorical variables are converted into numerical representations. Numerical features are scaled to a comparable range. Feature selection and feature reduction are also considered to retain important information and reduce redundant data. Data balancing is applied to maintain a suitable representation of PCOS and non-PCOS cases.

For ultrasound images, preprocessing includes noise reduction, contrast enhancement or grayscale conversion, resizing, normalization, and data augmentation. These operations prepare the images for CNN-based processing and improve the consistency of the image input.

C. CNN -Based Ultrasound Image Prediction

When an ultrasound image is available, the processed image is provided to the Convolutional Neural Network (CNN) module. The CNN automatically learns relevant visual patterns from the ultrasound images rather than requiring manual extraction of image features.

The model analyzes features related to ovarian structure and follicular patterns and uses the learned representations to classify the image as PCOS or non-PCOS. The CNN produces a prediction along with a probability score, which is passed to the hybrid fusion module.

The image prediction pipeline is:

Ultrasound Image → Image Preprocessing → CNN → Feature Learning → PCOS/Non-PCOS + Probability Score

D. SVM Prediction

After preprocessing, the clinical and lifestyle features are provided to the Support Vector Machine (SVM) model. The SVM learns the relationship between the processed health parameters and the corresponding PCOS and non-PCOS classes.

The model considers the available clinical and hormonal information and produces a prediction indicating whether the input is associated with PCOS or non-PCOS, together with a probability or confidence score.

The clinical prediction process can therefore be represented as:

Clinical Data → Data Preprocessing → Feature Processing → SVM → PCOS/Non-PCOS + Probability Score

E. Hybrid Prediction Fusion

The hybrid fusion module acts as the main decision-making stage when both clinical data and ultrasound images are available. The outputs generated by the SVM and CNN models are combined to obtain a final prediction.

It specifies possible fusion approaches including weighted

averaging, rule-based fusion, or a meta-classifier. The fusion process produces the final PCOS prediction together with an overall confidence score.

When an ultrasound image is unavailable, the system uses the SVM prediction obtained from the clinical data. This allows

the system to operate even when image information is not available.

The overall prediction process is therefore

SVM Output + CNN Output → Hybrid Fusion → Final PCOS Prediction + Confidence Score

F. Risk Assessment

Following the final prediction, the system performs risk assessment to classify the user's overall PCOS risk level.

The risk assessment considers the prediction result and other relevant health parameters such as BMI, age, stress, lifestyle factors, and available hormone information.

Based on these factors, the system categorizes the user into one of three risk levels:

Low Risk / Moderate Risk / High Risk

This module extends the system beyond simple PCOS/non-PCOS classification by providing an additional risk-level assessment.

G. Personalized Lifestyle Recommendation

The lifestyle recommendation module uses the user's prediction and risk level to generate personalized health suggestions.

Rather than providing identical recommendations to every user, the system considers individual health information and risk level.

The recommendations include:

- Diet plan
- Exercise suggestions
- Sleep-related guidance
- Stress-management techniques

These recommendations are intended to support preventive management and encourage healthier daily habits.

H. Report Generation

As shown in Fig 3.1, after prediction, risk assessment, and recommendation generation, the system creates a comprehensive report. The report contains the user details, final prediction, probability/confidence score, risk level, important contributing factors, and personalized lifestyle recommendations.

The generated report can be provided in PDF format, allowing the user to retain the results for future reference.

I. Database management and User Interface

The database module stores the user's profile, health information, prediction results, reports, and previous records. Maintaining these records enables users to access their previous results and monitor their information over time.

The user interface provides a dashboard through which users can enter their health information, upload ultrasound images, view prediction results, check their risk level, access lifestyle

recommendations, download reports, and view previous records.

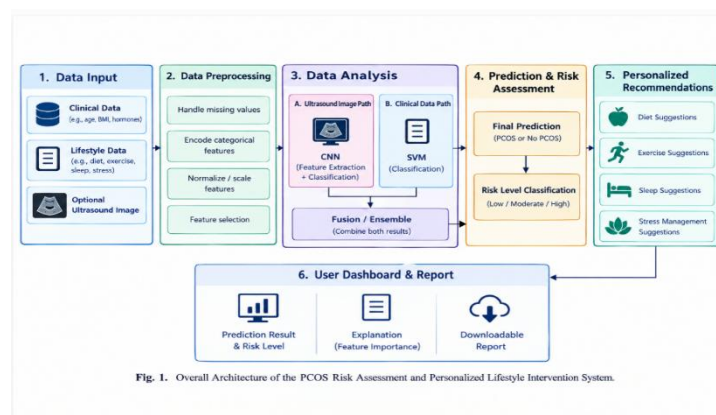


Fig. 1. Overall Architecture of the PCOS Risk Assessment and Personalized Lifestyle Intervention System.

Fig: 3.1. Architecture diagram

IV. RESULT ANALYSIS

A. Experimental Data

The proposed PCOS prediction system was evaluated using a clinical dataset of 1,500 patient records. The dataset includes clinical, demographic, lifestyle, and hormonal attributes used for tabular prediction. Of the 1,500 records, 1,255 patients had corresponding ultrasound images, which allowed the image-based CNN model to be trained and evaluated.

The tabular data included numerical features such as age, height, weight, cycle length, sleep hours, stress score, LH, FSH, insulin, and testosterone. Categorical features covered menstrual cycle characteristics, acne, hair growth, weight gain, family history, previous conditions, diet, and exercise. The preprocessing pipeline prepared the data for machine learning by handling the various numerical and categorical attributes prior to model training. This aligns with the proposed preprocessing module, which covers missing-value handling, feature encoding, feature scaling, feature selection, and data balancing.

Table 1. PCOS data summary and distribution

Parameter	Result
Clinical dataset size	1,500
Normal (No PCOS) cases	500
PCOS cases	1000
Patients with matched ultrasound images	1,255
Valid ultrasound images	1,255

The fine-tuning results show that the model maintained an accuracy 96.63%.

B. SVM model performance

The SVM model was trained on clinical and health-related tabular data and achieved a test accuracy of 99.53%.

It shows that the selected clinical, lifestyle, and hormonal features carry meaningful information for distinguishing between PCOS and non-PCOS cases within the evaluated dataset. The model works with numerical and categorical health attributes, including age, height, weight, menstrual cycle information, sleep, stress, hormone measurements, lifestyle factors, and medical history variables. Because of this, the system offers a tabular-data-based prediction pathway that can be used even when an ultrasound image is not available.

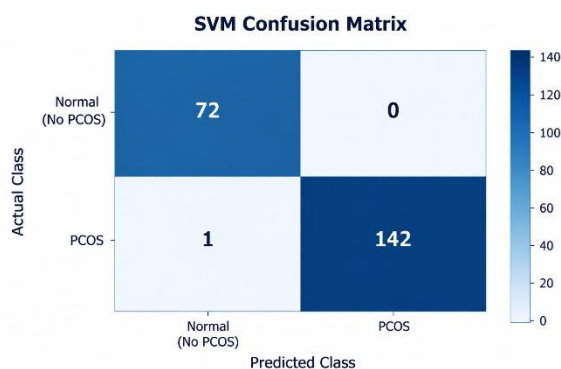


Fig. 4.1 Confusion matrix of the SVM clinical prediction model

Fig. 4.1 presents the confusion matrix of the Support Vector Machine (SVM) clinical prediction model for PCOS prediction. Of the 72 actual Normal (No PCOS) patients, all 72 were correctly classified as Normal, with no false-positive predictions. Among the 143 actual PCOS patients, 142 were correctly classified as PCOS, while one patient was incorrectly classified as Normal, resulting in a single false negative. As shown in Fig. 1, the model achieved an overall test accuracy of 99.53%, reflecting strong classification performance on the evaluated test dataset. The results also indicate a high capability to identify both Normal and PCOS cases within the test set. That said, further evaluation using independent patient data is needed to confirm the model's generalizability and reliability.

C. CNN Ultrasound model performance

The ultrasound component was trained using the 1,255 patients with matched ultrasound images. The available image dataset was divided into 1,004 training images and 251 validation images, with the images belonging to two classification classes.

The improvement in accuracy during this phase indicates that the CNN learned useful image features during training. At the end of the warm-up phase, the training accuracy reached high.

D. CNN Validation Performance

After training, the CNN was evaluated on the validation set. The model achieved a test accuracy of 96.63%.

The test accuracy of 96.63% demonstrates that the trained CNN was able to classify the ultrasound images with high accuracy on the validation data used during evaluation.

It should be noted that the reported CNN result is a test accuracy, whereas the reported SVM result is a test accuracy. Therefore, the two values should not be interpreted as a direct one-to-one comparison of model superiority because they

were obtained using different evaluation settings.

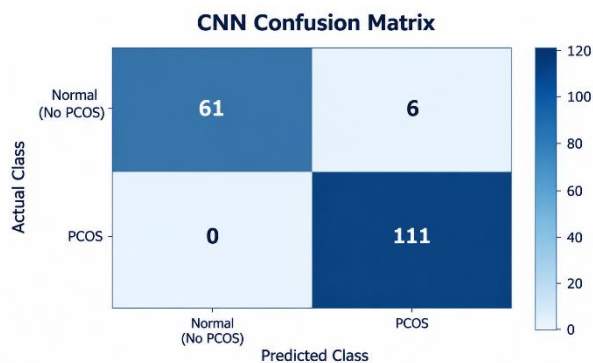


Fig 4.2. Confusion matrix of the CNN ultrasound classification model

Fig. 4.2 shows the confusion matrix of the Convolutional Neural Network (CNN) ultrasound classification model. Out of 67 actual Normal (No PCOS) test images, 61 were correctly classified as Normal, while six were incorrectly classified as PCOS, producing six false positives. All 111 actual PCOS test images were correctly classified as PCOS, with no false-negative predictions. As shown in Fig. 2, the CNN achieved a test accuracy of 96.63%, reflecting strong classification performance on the evaluated ultrasound images. These results indicate that the model successfully identified all PCOS cases in this test set, though a small number of Normal cases were misclassified. Additional testing on independent datasets is needed to determine whether this level of performance holds across different patient populations and imaging conditions.

E. Comparative Analysis of SVM and CNN

The two trained models achieved high classification accuracy on their respective evaluation datasets.

The SVM achieved a test accuracy of 99.53%, while the CNN achieved a test accuracy of 96.63%. The results demonstrate that both tabular and image-based approaches can be incorporated into the proposed PCOS prediction framework.

The SVM provides a clinical-data-based prediction pathway, while the CNN provides an image-based prediction pathway. This multimodal design is consistent with the proposed architecture, where the CNN analyzes ultrasound images and the SVM processes clinical and hormone-related information.

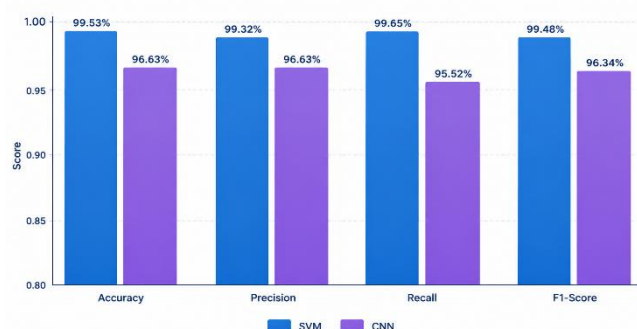


Fig. 4.3 SVM and CNN performance comparison

Fig. 3 compares the test accuracy of the SVM clinical

prediction model and the CNN ultrasound classification model. The SVM achieved an accuracy of 99.53% on 215 test patients, while the CNN achieved an accuracy of 96.63% on 178 test images. As shown in Fig. 3, the SVM accuracy was higher than that of the CNN by 2.90 percentage points. This difference suggests that the SVM performed better in terms of overall classification accuracy on its evaluated test set, while the CNN also showed high accuracy using ultrasound images. However, since the models were evaluated on different test-set sizes and input modalities, this difference alone cannot establish that the SVM is inherently superior. A fair comparison would require evaluating both models on the same eligible held-out patients with consistent ground-truth labels, as well as investigating potential data leakage before drawing any definitive conclusions.

F. Multimodal Prediction Framework

The proposed system is designed to combine information from clinical data and ultrasound images when both are available. The SVM generates a prediction from tabular health information, while the CNN generates a prediction from ultrasound images. The prediction fusion module is designed to combine these outputs and produce an overall prediction and confidence score.

In the current experimental results, the individual SVM and CNN models were successfully trained and saved. The SVM achieved 99.53% test accuracy, while the CNN achieved 96.63% test accuracy. These results establish the performance of the two individual components of the proposed multimodal framework.

G. System Implementation and Functional Validation

The trained models were successfully integrated into the developed application. The application successfully loaded the preprocessing pipeline, SVM model, and CNN model.

The system was executed using a Flask-based application running locally at port 5000. Prediction requests were successfully processed through the /predict endpoint. User login requests were also successfully processed, and generated reports were retrieved through the report endpoint.

The execution logs show multiple successful prediction requests with HTTP status 200, successful login requests with HTTP status 200, and successful report retrieval requests. This demonstrates that the trained models were not only evaluated independently but were also integrated into the working application.

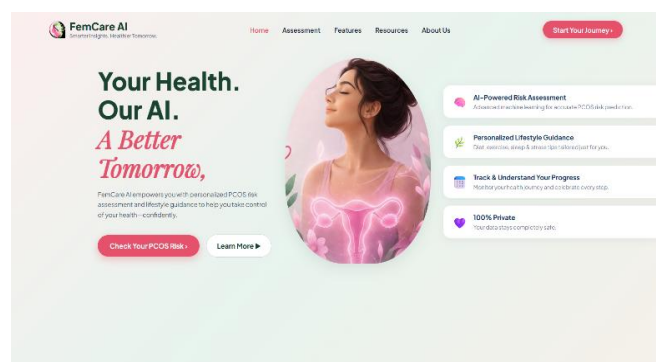


Fig.4.1 FemCare AI system home interface.

The Fig.4.1 depicts the main home interface of the proposed FemCare AI system. The interface introduces the system's key functionalities, including AI-powered PCOS risk assessment, personalized lifestyle guidance, health-progress tracking, and privacy-focused data handling. Fig.4.1 also highlights the system's primary navigation options and provides users with direct access to the PCOS risk assessment and additional information about the platform.

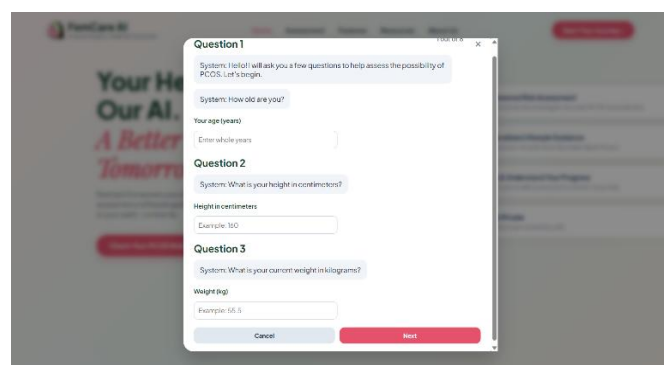


Fig.4.2 Interactive PCOS assessment questionnaire

Fig.4.2 shows the interactive PCOS assessment questionnaire used to collect user health and anthropometric information. As demonstrated in Fig.4.2, the system follows a conversational interaction format and collects inputs such as age, height, and weight through sequential questions. This interface is designed to make the assessment process simple and user-friendly while gathering the required inputs for PCOS risk prediction.

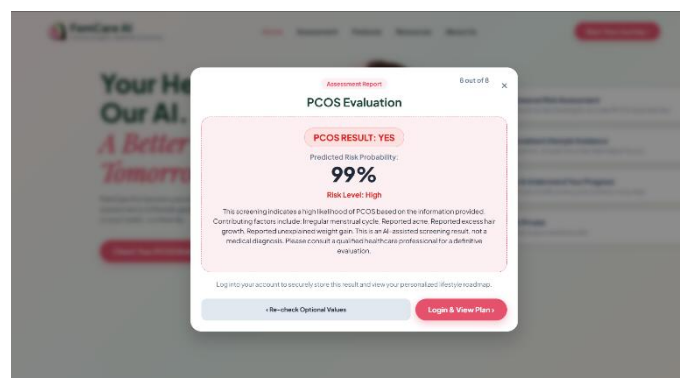


Fig.4.3 PCOS risk assessment result interface

Fig.4.3 presents the PCOS assessment result interface, where the system displays the predicted PCOS outcome along with the estimated risk probability and corresponding risk level. As shown in Fig.4.3, the interface also identifies the contributing factors associated with the assessment result and provides options for re-checking optional health parameters or proceeding to the personalized lifestyle plan.

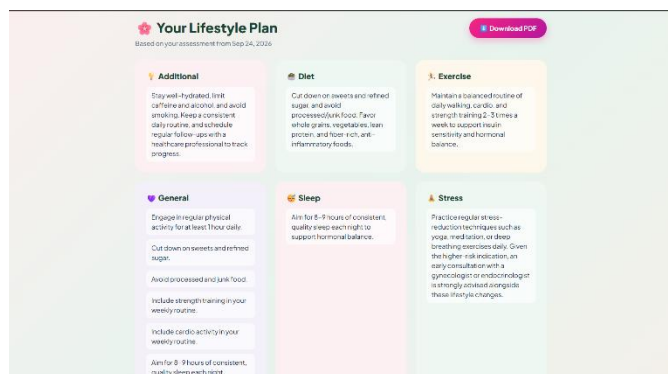


Fig.4.4 Personalized lifestyle recommendation interface

Fig.4.4 illustrates the personalized lifestyle recommendation interface of the proposed FemCare AI system. The interface presents recommendations under different categories, including diet, exercise, sleep, stress management, general health, and additional lifestyle practices. These recommendations are generated based on the user's PCOS assessment results and are intended to provide structured guidance for maintaining a healthier lifestyle.

V. CONCLUSION

The proposed Data-Driven Predictive Model for PCOS Risk Assessment and Personalized Lifestyle Intervention offers an integrated approach to supporting early PCOS risk identification and preventive healthcare. The system brings together clinical, lifestyle, psychological, and optional hormonal data with machine learning-based prediction. When an ultrasound image is available, the CNN module can analyze image-based ovarian patterns, while the SVM module handles tabular health information. The proposed hybrid framework combines these prediction outputs to generate a final PCOS prediction along with a confidence score. Beyond prediction, the system evaluates the user's overall risk as Low, Moderate, or High by taking into account factors such as BMI, symptoms, lifestyle, stress, age and any available hormone data. Personalized recommendations covering diet, exercise, sleep, and stress management are then provided based on the user's risk level and health profile. The system also supports report generation, data storage, history tracking and a user-friendly dashboard, making it more comprehensive than systems that offer only a PCOS classification. By bringing prediction, risk assessment, and personalized lifestyle guidance together within a single framework, the proposed system aims to support early awareness, preventive care and informed healthcare decision-making.

REFERENCES

- [1] Jayati Dixit, Anshuman Rai, Pulak Tandon, and Dhiraj Pandey, "ASHA: Machine Learning-Based Polycystic Ovary Syndrome (PCOS) Detection and Prediction System," Pune, India, Jan. 16–17, 2025, pp. 1–5, IEEE, 2025.
- [2] Xianqing Hu, Amit Yadav, Asif Khan, Abhishek Pratap Sah, and Sami Azam, "Construction of PCOS Prediction Model Based on BP Neural Network," pp. 885–889, 2025.
- [3] S. Singh, N. Pal, S. Shubham, D. K. Sarma, V. Verma, F. Marotta, and M. Kumar, "Polycystic Ovary Syndrome: Etiology, Current Management, and Future Therapeutics," *Journal of Clinical Medicine*, vol. 12, no. 4, p. 1454, 2023.
- [4] J. J. Benjamin, K. MaheshKumar, V. Radha, K. Rajamani, N. Puttaswamy, T. Koshy, K. N. Maruthy, and R. Padmavathi, "Stress and polycystic ovarian syndrome – a case control study among Indian women," *Clinical Epidemiology and Global Health*, vol. 22, p. 101326, 2023.
- [5] A. Sawant, N. Sayyad, S. Ugale, M. Gaikwad, and T. Kande, "PCOS: An Overview of Lifestyle Disorder Found in 90% Females Worldwide," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 10, no. 12, pp. C233–C239, Dec. 2022.
- [6] G. Konduri, A. Bollam, and R. Balaprabha, "Assessment of Anxiety, Stress and Depression in Women With and Without PCOS: A Comparative Study," *International Journal of Creative Research Thoughts*, vol. 10, no. 5, pp. d807–d811, May 2022.
- [7] I. Waghmare, "Effect of Polycystic Ovary Syndrome (PCOS) in Indian Women of Reproductive Age", *International Journal of Creative Research Thoughts*, vol. 10, no. 1, Jan. 2022.
- [8] A. G. Wale, S. S. Walvekar, and S. R. Zine, "Female Infertility: A Review on Definition, Causes and Its Treatment," *International Journal of Creative Research Thoughts*, vol. 8, no. 9, Sep. 2020.
- [9] Priyanka Nalawade, Vishal Borate, Rajlakshmi N. Shindalkar, Snehal S. Bansode, Shreya D. More, and Cynthia Pachal, "Explainable CNN & Hybrid Deep Learning Model for Health Risk PCOS Prediction & Diagnosis," *International Conference on Intelligent Embedded, Microelectronics, Communication and Optical Networks*, 2025.