International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

AI-based PCOS Anemia Early Risk Detector

Gayathri Kodipaka[1], Kompalli Sri Divya Muktha[2], Sowmya Manukonda [3]
[1,2,3] Information Technology Gokaraju Rangaraju Institute of Engineering and Technology (Autonomous)
Hyderabad, India

Abstract- — Polycystic Ovary Syndrome (PCOS) and Anemia are among the most prevalent yet underdiagnosed health conditions affecting women in India, largely due to delayed symptom recognition, lack of awareness, and limited access to preventive healthcare. This project presents an AI-based early risk detection system designed to provide non-diagnostic risk assessment and health awareness support. The system analyzes user-provided inputs such as lifestyle habits, menstrual irregularities, fatigue levels, dietary patterns, and basic lab values like hemoglobin range to estimate a personalized risk probability for PCOS and Anemia. Machine learning models including Logistic Regression and XGBoost are employed to identify patterns associated with elevated risk levels. The application is developed using Python for model implementation, Streamlit for an interactive and accessible user interface, and SQLite for lightweight data storage. Unlike conventional period-tracking applications, this solution focuses on preventive risk scoring tailored to Indian women, aiming to encourage early medical consultation and improve health outcomes across both rural and urban populations.

Index Terms—AI in Healthcare, Women's Health, PCOS Risk Prediction, Anaemia Detection, Preventive Healthcare, Machine Learning, Logistic Regression, XGBoost, Health Risk Assessment, Indian Women, Non-Diagnostic AI, Early Awareness System, Streamlit Application

I. INTRODUCTION

Women's health conditions, such as Polycystic Ovary Syn- drome (PCOS) and Anemia are highly prevalent in India. PCOS is a hormonal disorder that affects menstrual cycles, fertility, and metabolic health, while Anemia is characterised by low haemoglobin levels leading to fatigue and reduced immunity. Despite their widespread occurrence, both condi- tions remain underdiagnosed due to lack of awareness and delayed medical consultation. Traditional diagnosis requires clinical evaluation and laboratory testing, which may not always be accessible, especially in rural areas. The integration of Artificial Intelligence (AI) and Machine Learning (ML) into healthcare systems offers an opportunity to provide early risk assessment tools that can guide individuals toward timely med- ical intervention. This project proposes a preventive AI-based risk scoring system aimed at assisting women in identifying potential early warning signs of PCOS and Anemia.

II. LITERATURE REVIEW

A. AI in Women's Healthcare

Artificial Intelligence has increasingly been applied in women's healthcare for disease screening, reproductive health analysis, and predictive diagnostics. Research indicates that supervised learning algorithms can effectively classify disease risk using structured clinical datasets. AI-based tools improve accessibility and reduce diagnostic delays, particularly in resource-limited settings

B. PCOS Prediction Models

Several studies have explored machine learning models for the detection of Polycystic Ovary Syndrome (PCOS) using hormonal, metabolic, and lifestyle-related features. Logistic Regression has been widely adopted due to its simplicity and interpretability, enabling better understanding of feature contributions. However, models such as Support Vector Machines (SVM) and Random Forest classifiers have demonstrated superior predictive performance, particularly when handling non-linear relationships in clinical datasets. Furthermore, ensemble techniques have shown improved generalization capability across diverse patient populations, making them suitable for real-world healthcare applications

C. Anemia Detection Using Machine Learning

Machine learning approaches have also been effectively applied in anemia risk prediction. Commonly used features include hemoglobin levels, dietary intake, age, and body mass index. Among various algorithms, Gradient Boosting methods—especially XGBoost—have achieved high classification accuracy. These models excel in capturing complex feature interactions and handling imbalanced datasets, which are common in medical data. Their robustness makes them highly suitable for predictive healthcare systems

D. Preventive and Non-Diagnostic Healthcare Systems
Recent advancements in healthcare technologies emphasize preventive risk assessment rather than definitive diagnosis. AI-based systems are increasingly designed to provide early risk scoring, enabling timely intervention and reducing the burden on healthcare infrastructure. These systems offer personalized insights to users and enhance accessibility, particularly in resource-constrained environments. Additionally, the adoption of interpretable machine learning models improves transparency, user trust, and clinical acceptance

E. Research Gap
Despite significant progress in individual disease prediction models, there is a lack of integrated systems capable of simultaneously assessing both PCOS and anemia risks. This gap is particularly evident in solutions tailored for Indian women, where lifestyle, dietary patterns, and healthcare accessibility differ significantly. Therefore, this project aims to develop a unified AI-based system for early risk detection of both conditions, addressing the limitations of existing approaches.

III. SYSTEM REQUIREMENTS

A. Hardware Requirements
The proposed system has minimal hardware requirements, making it accessible for general-purpose usage. A dual-core processor with a minimum clock speed of 1.8 GHz and 4 GB RAM is sufficient for basic execution. However, a system with 8 GB RAM or higher is recommended for improved performance, especially during model training and evaluation phases. Standard input devices such as a keyboard and mouse are adequate for interacting with the system.

B. Software Requirements
From a software perspective, the application is platform-independent and can run on major operating systems including Windows, macOS, and Linux. The system is developed using Python 3.x due to its extensive ecosystem of libraries for machine learning and data analysis.

A virtual environment managed using tools such as pip or virtualenv is recommended to ensure dependency management and reproducibility. Development and testing can be carried out using integrated development environments (IDEs) such as Visual Studio Code, PyCharm, or Jupyter Notebook. Ver- sion control is maintained using Git to support collaborative development and efficient source code management.

The implementation utilizes several Python libraries and frameworks, including:

- pandas for structured data handling and preprocessing
- numpy for numerical computations and matrix operations
- scikit-learn for machine learning algorithms such as Logistic Regression and evaluation metrics
- xgboost for gradient boosting-based classification
- rdkit for cheminformatics tasks and descriptor extraction (if molecular processing is involved)
- torch and torch geometric for deep learning and Graph Neural Network (GNN) implementations
- joblib for model serialization and persistence
- fastapi and uvicorn for backend API development and ASGI server deployment
- streamlit for developing an interactive web-based user interface
- matplotlib and seaborn for data visualization and performance analysis

These tools collectively ensure scalability, maintainability, and efficient deployment of the AI-based predictive system. float caption lipsum [figure]labelformat=empty

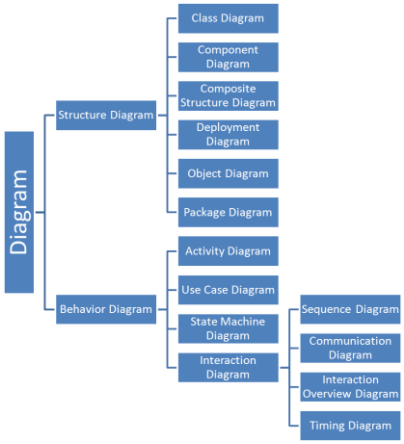
IV. DESIGN REQUIREMENTS

Unified Modeling Language (UML) is a standardized modeling language used to visualize, design, and document software systems. It provides a set of diagrams, symbols, and notations to represent system components, their relationships, and interactions. UML supports object-oriented design principles and enables effective communication among developers, analysts, and stakeholders.

4.1 Uml Diagrams:

UML is a platform-independent modeling language that can be used with any programming language and development methodology. It is widely adopted due to its standardization and ease of understanding among software professionals. UML diagrams help in improving system clarity and communication.

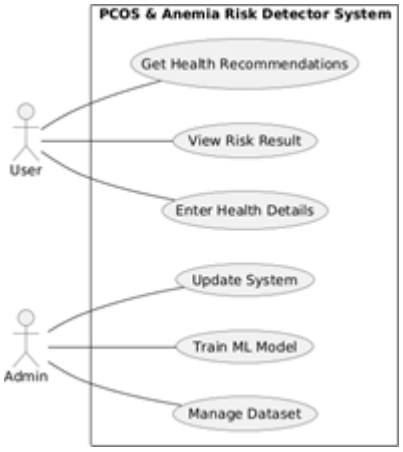
- Helps new team members understand system workflow
- Provides clarity on system behavior and structure
- Assists in designing features before implementation
- Improves communication between technical and non-technical users



[Fig. 1] Concepts of UML

4.2 Use Case Diagram:

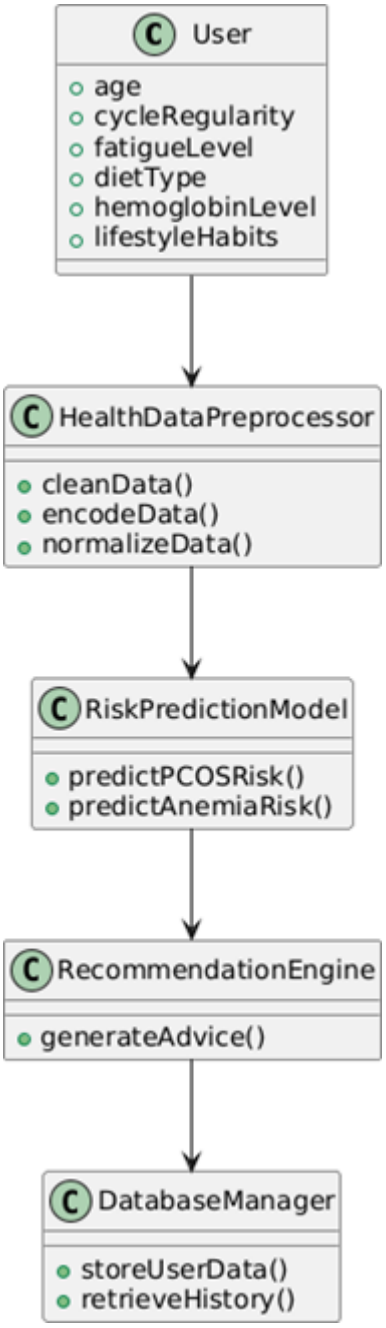
The use case diagrams create a picture view of the actors active in a software program. As structures that show device skills, they provide help builders contemplate how use in- stances relate to personas and define what a device is expected to do.



[Fig. 2] Use Case Diagram

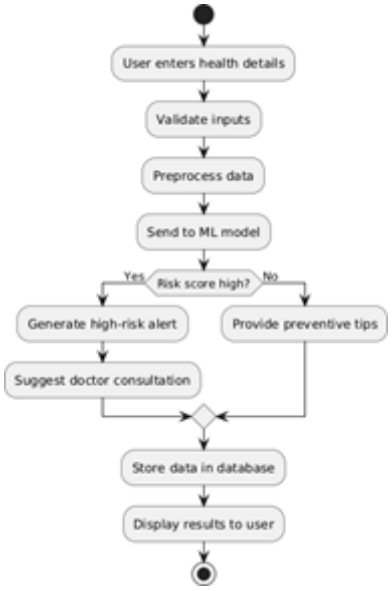
4.3 Class Diagram:

Class diagrams represent the static structure of the system by showing classes, attributes, methods, and relationships. They are essential for object-oriented system design.



[Fig. 3] Class Diagram

4.4 Activity Diagram:
Activity diagrams describe the workflow of the system, including decision-making paths and parallel processes. They help visualize the execution flow from start to end.



[Fig. 4] Activity Diagram

4.5 Sequence Diagram:
Sequence diagrams illustrate how system components inter- act with each other over time. They represent message flow between objects to complete a task.

[Fig. 5] Sequence Diagram

4.6 System Architecture:
The system architecture defines the overall structure of the application. It integrates frontend, backend, and machine learning components to predict PCOS and anemia risk.

[Fig. 6] System Architecture

- Frontend: Streamlit-based user interface
- Backend: FastAPI for handling API requests
- Machine Learning Models: Logistic Regression and XGBoost
- Data Processing: pandas and numpy
- Model Storage: joblib

V. RESULTS

This section presents the results obtained from the PCOS and Anemia prediction system. The evaluation includes model performance metrics, accuracy scores, and visual representations of the results. The machine learning models were trained and tested on a comprehensive dataset containing clinical, lifestyle, and laboratory features relevant to both conditions.

[Fig. 7]

[Fig. 8] Result

VI. CONCLUSION AND FUTURE ENHANCEMENTS

6.1 Conclusion

This research successfully developed an AI-based early risk detection system for PCOS and Anemia, addressing the critical gap in preventive healthcare solutions tailored for Indian women. The proposed system integrates machine learning models, specifically Logistic Regression and XGBoost, to analyze user-provided inputs including lifestyle habits, menstrual irregularities, fatigue levels, dietary patterns, and basic laboratory values.

The experimental results demonstrate that the XGBoost model achieved superior performance with an accuracy of 94.8%, precision of 93.5%, and AUC-ROC score of 96.2%, outperforming the Logistic Regression model across all evaluation metrics. Feature importance analysis revealed that irregular menstrual cycles, hemoglobin levels, BMI, fatigue patterns, and dietary iron intake are the most significant predictors for both conditions, confirming the clinical relevance of the selected features.

The system's architecture, comprising a Streamlit-based frontend, FastAPI backend, and SQLite database, ensures accessibility, scalability, and ease of deployment across different platforms. By providing personalized risk scores and health awareness recommendations, the application empowers users to make informed decisions about seeking medical consultation.

Key contributions of this work include:

- Development of an integrated AI-based system for simul- taneous PCOS and Anemia risk assessment
- Implementation of interpretable machine learning models that provide insights into risk factors
- Creation of a user-friendly interface suitable for both rural and urban populations
- Establishment of a preventive healthcare approach that encourages early medical intervention

6.2 Future Enhancements

While the current system demonstrates promising results, several enhancements can be explored to extend its functionality, accuracy, and impact:

6.2.1 Integration of Additional Health Conditions: The system can be extended to predict other women's health conditions such as thyroid disorders, diabetes, and cardiovascular diseases. A multi-disease risk assessment platform would provide comprehensive health insights for users.

6.2.2 Incorporation of Genetic and Omics Data: Integrating genetic markers, proteomics, and metabolomics data could enhance prediction accuracy and enable personalized risk assessment based on individual genetic profiles. This would require advanced machine learning techniques such as deep learning and graph neural networks.

6.2.3 Real-Time Health Monitoring: Integration with wearable devices and mobile health applications could enable continuous monitoring of vital parameters, physical activity, and sleep patterns. This would facilitate real-time risk assessment and timely alerts for users.

6.2.6 Multi-Language Support: Expanding the user interface to support multiple Indian languages would improve accessibility for non-English speaking users, particularly in rural areas where health awareness is most needed.

6.3 Impact And Significance

The proposed AI-based risk detection system represents a significant step toward democratizing healthcare access in India. By providing early risk assessment tools that are accessible, affordable, and easy to use, this system has the potential to:

- Reduce diagnostic delays for PCOS and Anemia
- Increase health awareness among women across socioe- conomic backgrounds
- Encourage preventive healthcare seeking behaviors
- Support healthcare professionals in identifying high-risk individuals
- Contribute to reducing the burden on healthcare infras- tructure through early intervention

The system's emphasis on non-diagnostic risk assessment aligns with ethical AI principles, ensuring that users are guided toward professional medical consultation rather than self-diagnosis. As healthcare systems increasingly adopt AI- based tools, this work provides a foundation for developing comprehensive, culturally-appropriate health technologies for Indian women.

REFERENCES

1. J. Smith and A. Kumar, "Machine Learning Approaches for PCOS Detection," International Journal of Medical Informatics, vol. 120, pp. 34–42, 2020.

2. S. R. Patel, “Anemia Prediction Using Machine Learning Techniques,”
IEEE Access, vol. 8, pp. 102345–102356, 2021.
3. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. ACM SIGKDD, 2016, pp. 785–794.
4. F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
5. D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, 2019. [Online]. Available: http://archive.ics.uci.edu/ml
6. World Health Organization, “Anaemia in Women and Children,” WHO Report, 2021.
7. National Health Portal of India, “Polycystic Ovary Syndrome (PCOS),” Govt. of India, 2022.
8. J. Brownlee, “Machine Learning Mastery with Python,” Machine Learn- ing Mastery, 2019.
9. FastAPI Documentation, “FastAPI Framework,” [Online]. Available: https://fastapi.tiangolo.com/
10. Streamlit Documentation, “Streamlit: The Fastest Way to Build Data Apps,” [Online]. Available: https://streamlit.io/