AI – Powered PCOD Detection Platform

Prasad Vadkar *1, **Mrudula Mane***2, **Madhuri Kolekar***3, **Sanjana Jadhav***4
*1 Prasad Hanmant Vadkar, Computer Science , JCEP K. M. Gad, Karad ,Maharashtra,
*2Mrudula Lalaso Mane, Computer Science , JCEP K. M. Gad, Karad , Maharashtra,
*3Madhuri Kumar Kolekar, Computer Science, JCEP K. M. Gad, Karad, Maharashtra,
*4 Sanjana Arvind Jadhav, Computer Science, JCEP K.M. Gad, Karad, Maharashtra (A.N. Pawar, Computer Science, JCEP K. M. Gad, Karad, Maharashtra, India)
JAYAWANT COLLEGE OF ENGINEERING AND POLYTECHNIC , KILLEMACHINDRAGAD,SANGLI
AFFILIATED TO DR.BABASAHEB AMBEDKAR TECHNICAL UNIVERSITY, LONERE 2024-2025

Abstract

Polycystic Ovary Syndrome (PCOS), also referred to as Polycystic Ovarian Disease (PCOD), is one of the most prevalent endocrine disorders affecting women of reproductive age worldwide. It is a leading cause of anovulatory infertility and is characterized by hormonal imbalances that result in symptoms such as irregular menstrual cycles, excessive weight gain, acne, hair loss, and skin darkening. Despite its high prevalence, early-stage detection and accurate prediction of PCOS remain challenging due to limitations in existing diagnostic methods and treatment strategies.

This research aims to address these challenges by developing an advanced, computer-aided detection system utilizing machine learning (ML) and deep learning (DL) techniques. The system leverages ovary ultrasound (USG) images— one of the most reliable diagnostic modalities for PCOS—and incorporates a Convolutional Neural Network (CNN) for robust feature extraction. To enhance classification performance, a stacking ensemble model is implemented using a combination of traditional machine learning classifiers as base learners and bagging or boosting techniques as meta-learners. The CNN architecture is further strengthened through transfer learning and modern feature selection techniques such as I-SQUARE and CHI-square.

The study involves training and evaluating the proposed model on a dataset comprising 4000 ovary USG images, sourced from a publicly available PCOS dataset on Kaggle by Parson Kottarathil. Additionally, five ML classifiers—Random Forest, Support Vector Machine (SVM), Logistic Regression, Gaussian Naïve Bayes, and K-Nearest Neighbors—were evaluated on a subset of the dataset containing 41 clinical and physiological features, with the top 30 features selected for classification.

Experimental results indicate that the Random Forest Classifier outperforms other models in terms of accuracy and reliability. The proposed hybrid system significantly improves detection accuracy while reducing execution time, making it a promising solution for aiding healthcare professionals in the early diagnosis and management of PCOS.

This research lays the foundation for intelligent and scalable PCOS detection systems that integrate clinical data and medical imaging, thereby advancing personalized and timely healthcare delivery for women suffering from this condition.

Keywords:

Polycystic Ovary Syndrome (PCOS), Machine Learning, Deep Learning, Convolutional Neural Network (CNN), Medical Imaging, Ultrasound, Classification, Data Mining, Healthcare, Prediction System, Early Diagnosis

1. Introduction

Over the last few decades, rapid advancements in technology have brought significant changes to every aspect of human life, making daily tasks more efficient and improving quality of life. Among these advancements, Machine Learning (ML) has emerged as a powerful tool, particularly in the healthcare sector, where it offers the potential to automate diagnostics, process large datasets, and deliver predictive insights that aid clinical decision-making.

One such area where ML can make a notable impact is in the diagnosis and management of Polycystic Ovary Disease (PCOD)—a complex hormonal disorder that affects women during their reproductive years. PCOD is characterized by the presence of small cysts in the ovaries, hormonal imbalances (especially elevated androgens), and irregular ovulation. As a result, women with PCOD often experience irregular periods, acne, weight gain, excessive hair growth, and may face difficulties in conceiving. Despite being a common condition, around 70% of women with PCOD go undiagnosed, as highlighted in studies like [Dewailly, 2013].

The exact cause of PCOD is still not clearly understood, though it is believed to have genetic components. The condition remains unpredictable, and without clear trends or patterns, it becomes difficult to detect and treat effectively in its early stages. The high cost and time investment required for repeated clinical tests and scans also add burden to both patients and healthcare professionals.

To address this, there is a growing need for automated, intelligent systems that can assist in early diagnosis using easily available data. Machine Learning can analyze large sets of clinical parameters such as Follicle-Stimulating Hormone (FSH) levels, ovary size, and hormone profiles, and can be trained to detect patterns that indicate the presence of PCOD. Early detection not only allows timely treatment but also helps in preventing long-term health risks such as type-2 diabetes, cardiovascular diseases, and psychological stress.

This research aims to develop a PCOD prediction model using Machine Learning techniques, which can analyze both medical records and ovary ultrasound (USG) images to provide fast, reliable, and cost-effective diagnostic support. The proposed system will contribute toward bridging the gap between traditional diagnosis and intelligent healthcare, ultimately leading to better outcomes for women's health.

2. Objectives

The proposed PCOD detection system is developed with a set of clear, focused, and practical objectives aimed at improving early diagnosis and patient care using modern technological advancements. These objectives reflect the need for a smart, accessible, and efficient healthcare solution and guide the research and implementation of this system. The main objectives are:

1. **Early Diagnosis of PCOD**: Leverage machine learning algorithms to detect PCOD at an early stage using clinical data and ultrasound (USG) imaging, reducing the risk of long-term health complications.
2. **Data-Driven Insights**: Utilize large datasets to analyze trends and patterns in PCOD symptoms and hormone levels, offering more accurate predictions compared to traditional diagnosis methods.
3. **User-Friendly Diagnostic Platform**: Develop an intuitive, easy-to-use interface for healthcare providers and patients to interact with the diagnostic system and view results with clarity.
4. **Reduction in Healthcare Costs**: Minimize the number of clinical tests and imaging scans by using intelligent prediction models, making PCOD diagnosis more affordable for patients.
5. **Enhancement of Clinical Support**: Provide doctors and healthcare workers with a reliable support system that assists in decision-making and reduces manual diagnostic errors.
6. **Awareness and Education**: Spread awareness about PCOD symptoms, causes, and lifestyle changes by integrating informative content and health guidelines within the platform.
7. **Performance Accuracy**: Ensure high diagnostic accuracy through the implementation of advanced models such as Convolutional Neural Networks (CNN) for image analysis and ensemble classifiers for predictive analytics.
8. **Scalability and Flexibility**: Design the system to be scalable across different healthcare settings—from clinics to large hospitals—ensuring adaptability and ease of integration.
9. **Privacy and Security**: Safeguard patient data with strict data protection measures to ensure confidentiality and compliance with medical data regulations.
10. **Future Integration with Healthcare Systems**: Build a flexible framework that can be integrated with existing Electronic Health Records (EHR) and hospital management systems for wider adoption.

3. Methodology

2.1 Algorithm Used:

The PCOS Detection and Prediction System has been developed following a structured and iterative methodology that integrates both data science practices and modern software development principles. The methodology ensures that the system is accurate, efficient, scalable, and user-friendly, aimed at assisting early diagnosis of PCOS in women. The major phases of the methodology are as follows:

1. **Requirement Analysis:** The project commenced with detailed research into PCOS, collecting requirements from medical professionals, academic literature, and existing PCOS diagnosis practices. This phase focused on understanding user needs, medical parameters, data availability, and diagnostic challenges in early detection.

2. **Data Collection and Preparation:** Relevant data was collected from publicly available PCOS datasets, including ultrasound images and clinical features such as hormone levels, BMI, insulin resistance, and menstrual irregularity. Data preprocessing included:

- Handling missing values
- Feature selection (using Chi-square test)
- Image augmentation (resizing, flipping, zooming)
- Normalization of numerical inputs

3. **System Design and Architecture Planning:** A Based on the analysis, the system architecture was designed using Python-based technologies and machine learning frameworks. The design included data flow diagrams, algorithm selection strategy, and model deployment flow. Emphasis was placed on user- friendly interface design, scalable backend structure, and seamless integration between modules.

4. **Algorithm Selection and Prototyping:** Prototypes were developed to evaluate different algorithms using smaller data samples. Machine learning models such as Random Forest, SVM, Logistic Regression, and CNN (Convolutional Neural Network) were tested. This stage helped select the most efficient algorithms based on accuracy and execution speed.

5. **Development Iterations:** The Using an iterative development approach:

- Clinical feature models and USG image models were developed in parallel.
- CNN was implemented for ultrasound image feature extraction.
- Ensemble machine learning was applied (stacking with base and meta learners).
- Python (TensorFlow/Keras and Scikit-learn) was used as the main development environment.

6. **Frontend and Back-end Development:**

- **Frontend**: A basic GUI or web interface was built using Python libraries or Flask to allow user interaction and input handling.
- **Backend**: ML models were integrated to process clinical inputs or images and return predictions. Data management, image processing, and result display were handled in the backend.

7. **Testing and Validation :** Extensive testing was conducted to ensure system reliability:

- Unit testing of individual models and modules
- Integration testing for complete workflow
- Performance testing to evaluate accuracy, recall, precision, and F1-score

- User acceptance testing (UAT) to validate usability from a medical user perspective

8. **Deployment and Future Enhancements:** The final trained model was saved and packaged for deployment using .h5 format (for CNN) and .pkl for ML classifiers. Future scope includes:

- Integration into hospital systems
- Mobile app version for wider accessibility
- Real-time patient monitoring and predictive alerts

4. Process of PCOS Detection System

1 Process of PCOS Detection System:

he PCOS Detection and Prediction System follows a structured, step-by-step process to ensure accurate diagnosis using machine learning and deep learning techniques. This system is designed to work efficiently from data collection to final prediction, making it suitable for real-time use by healthcare professionals and patients.

Five steps process

Step 1: Data Collection and Understanding

The process begins with gathering a reliable dataset, including clinical parameters (such as BMI, insulin levels, hormone values) and ultrasound images of ovaries. The dataset used in this project was obtained from Kaggle, contributed by Parson Kottarathil, consisting of both PCOS and non-PCOS patient data.

Step 2: Data Preprocessing and Feature Selection

Once data is collected, it undergoes cleaning and transformation. Steps include:

- Handling missing values
- Encoding categorical values
- Normalizing numerical fields
- Selecting the top 30 most relevant features using Chi-Square and I-SQUARE methods

This step is crucial to reduce noise and improve model efficiency.

Step 3: Model Training and CNN-based Image Analysis

After preprocessing, different machine learning classifiers are trained using the cleaned dataset. These include:

- Random Forest
- Support Vector Machine (SVM)
- Logistic Regression
- Gaussian Naïve Bayes
- K-Nearest Neighbors

Meanwhile, Convolutional Neural Networks (CNNs) are used to extract features from ovary ultrasound images, enhancing diagnosis accuracy through image-based learning.

Step 4: Ensemble Learning for Enhanced Prediction

To further boost accuracy, the system uses a Stacking Ensemble approach where:

- The trained classifiers serve as base learners
- A bagging or boosting model functions as the meta-learner

This method combines the strengths of multiple models, reducing bias and variance and improving the reliability of final predictions.

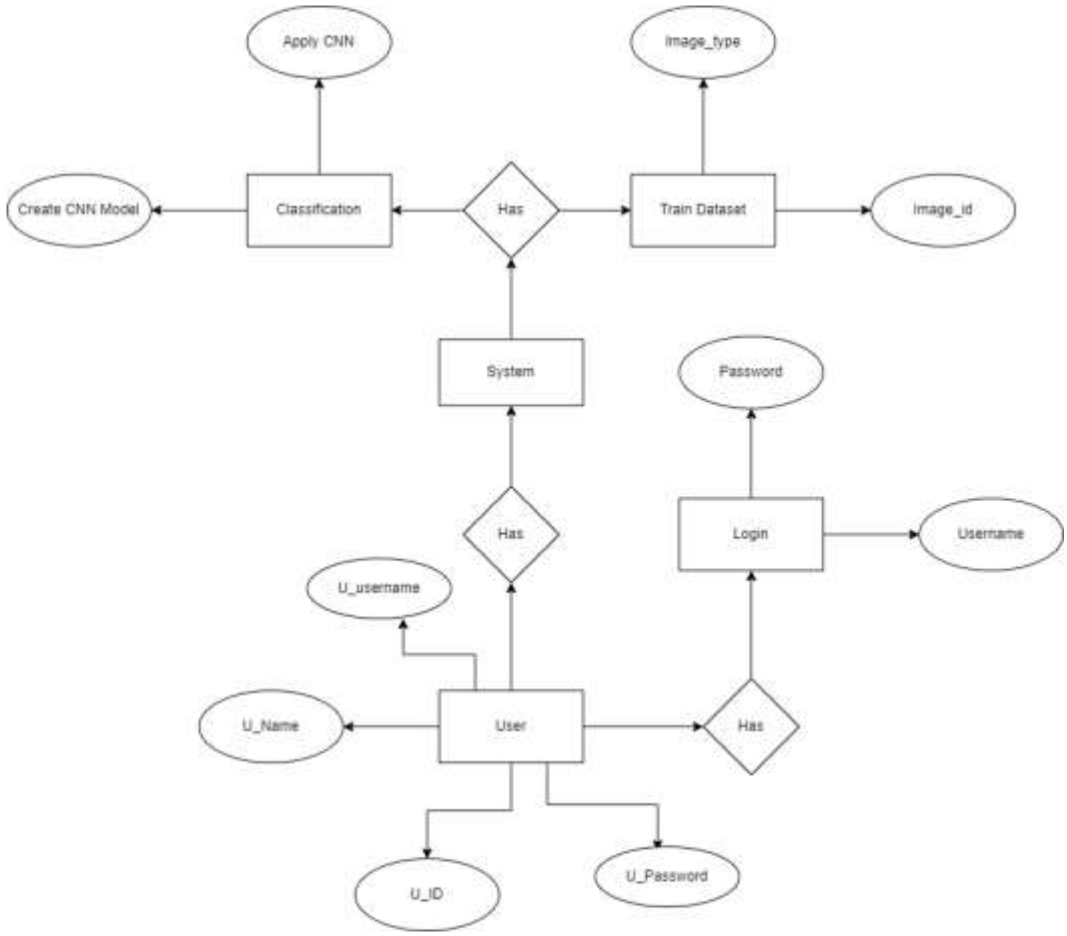
Step 5: Evaluation and Result Interpretation

The final model is tested using performance metrics like:

- Accuracy
- Precision
- Recall

Results are displayed through a user-friendly interface or visualizations, allowing healthcare providers to interpret whether a patient is likely to have PCOS and what factors contribute most to that prediction.

1. Flow of Execution



5. Conclusion

The purpose of the **PCOS Detection and Prediction** project is rooted in the vision of transforming women's healthcare through the integration of machine learning technologies and clinical intelligence. This project addresses the growing need for timely, accessible, and accurate diagnosis of PCOS—a condition affecting millions of women globally—by leveraging advanced data-driven methods.

- **Empowering Early Diagnosis:** At its core, the project aims to support early detection of PCOS through intelligent algorithms that analyze both clinical and ultrasound data. By reducing dependence on multiple costly tests, it empowers both patients and healthcare providers with quick and actionable insights.
- **Accessibility and Scalable Technology:** The proposed system provides a user-friendly and scalable platform that can be deployed in clinics, hospitals, or even used directly by patients. This accessibility helps bridge the gap in women's healthcare, especially in under-resourced or rural areas.
- **Supporting Healthcare Professionals:** The project enhances the diagnostic capabilities of doctors by offering reliable predictions through ensemble machine learning and CNN-based imaging systems. It acts as a decision-support tool, improving clinical workflow and reducing the burden of manual interpretation.
- **User-Centric Healthcare Experience:** With an intuitive design and seamless functionality, the system is built with a focus on usability, ensuring that both tech-savvy users and healthcare workers can operate it with ease. This improves engagement and fosters trust in technology-aided diagnosis.
- **Data-Driven Innovation in Medicine:** This project exemplifies how medical diagnostics can benefit from the latest innovations in data science. By employing CNN, Random Forest, SVM, and ensemble learning, it delivers high accuracy and precision, supporting better outcomes for women with PCOS.
- **Commitment to a Healthier Future:** In line with broader goals of wellness and preventive care, the PCOS prediction system encourages timely lifestyle changes and ongoing monitoring. Its implementation contributes toward reducing the long-term health risks associated with untreated PCOS—such as diabetes, cardiovascular disorders, and infertility.

6. Acknowledgement

We would like to express our sincere gratitude to all individuals and institutions who supported and guided us throughout the development of this research project on PCOD Detection and Prediction using Machine Learning.

First and foremost, we would like to thank Prof. A. N. Pawar for his valuable mentorship, encouragement, and continuous guidance. His insightful feedback and expert knowledge played a significant role in shaping the direction, scope, and technical depth of this project.

We extend our heartfelt thanks to our college management and faculty members at Jaywant College Of Engineering And Polytechnic, whose support and resources enabled us to conduct our research efficiently. We are also grateful to the Department of Computer Engineering for providing the infrastructure and environment needed for experimentation and learning.

Special appreciation goes to our team members — Madhuri Kolekar, Mrudula Mane, Prasad Vadkar, and Sanjana Jadhav — for their collaboration, dedication, and teamwork during all phases of this project. Each member's contribution to data collection, preprocessing, model building, and report writing was invaluable to the successful completion of this study.

7. Authors' Biography

- **Madhuri Kumar Kolekar** **- Student**
- **Mrudula Lalaso Mane** **-Student**
- **Prasad Hanmant Vadkar** **- Student**
- **Sanjana Arvind Javdhav** **- Student**

8. References

Example of List of References

1] Palak Mehrotra, Jyotirmoy, Chatterjee, Chandan Chakraborty, "Automated Screening of Polycystic OSyndrome using Machine Learning Techniques", IEEE, 2012.

[2] Bedy Purnama, Untari Novia Wisesti, Adiwijaya, Fhira Nhita, Andini Gayatri, Titik Mutiah, "A Classification of Polycystic Ovary Syndrome Based on Follicle Detec-tion of Ultrasound Images, 2015 3rd International Conference on Information and Communication Tech- nology (ICoICT).

[3] Amsy Denny, Anita Raj, Ashi Ashok, Maneesh Ram C, Remya George, "i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOD) Using Machine Learning Techniques", 2019 IEEE Region 10 Conference (TENCON 2019).

[4] Subrato Bharati, Prajoy Podder, M. Rubaiyat Hossain Mondal, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms". 2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh.

[5] Ning-Ning Xie, Fang-Fang Wang, Jue Zhou, Chang Liu, Fan Qu, "Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network", Hindawi BioMed Research International Volume 2020.

6] Priyanka R. Lele, Anuradha D. Thakare, "Comparative Analysis of Classifiers for Polycystic Ovary Syndrome Detection using Various Statistical Measures", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181: Vol. 9 Issue 03, March-2020.

[7] Namrata Tanwani, "Detecting PCOD using Machine Learning", IJMTES | International Journal of Modern Trends in Engineering and Science ISSN: 2348-3121, Volume:07 Issue:01 2020.

[8] J. Madhumitha, M. Kalaiyarasi, S. Sakthiya Ram, "Automated Polycystic Ovarian Syndrome Identification with Follicle

[7] Namrata Tanwani, "Detecting PCOD using Machine Learning", IJMTES | International Journal of Modern Trends in Engineering and Science ISSN: 2348-3121, Volume:07 Issue:01 2020

[8] J. Madhumitha, M. Kalaiyarasi, S. Sakthiya Ram, "Automated Polycystic Ovarian Syndrome Identification with Follicle
Recognition", 2021 3rd International Conference on Signal Processing and Communication

[9] Pijush Dutta, Shobhandeb Paul, Madhurima Majumder, "An Efficient SMOTE Based Machine Learning classification for Prediction & Detection of PCOD", Research Square, November 8th, 2021.

[10] Muhammad Sakib Khan Inan, Rubaiath E Ulfath, Fahim Irfan Alam, Fateha Khanam Bappee, Rizwan Hasan, "Improved Sampling and Feature Selection to Support Extreme Gradient Boosting for PCOD Diagno