WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Empowering Women's Health: Machine Learning for PCOS Detection and Prediction

J. NISHA[1], S. PAVITHRA[1], S. SAKTHI[1], P. SATHYA[1], Dr. S. BALAJI[2], M. SAMUNDEESWARI[3]

[1]UG Scholars, Department of CSE, Kingston Engineering College, Vellore
[2] Assistant Professor, Department of CSE, Kingston Engineering College, Vellore
[3] Assistant Professor, Department of CSE, Kingston Engineering College, Vellore

ABSTRACT :

Polycystic Ovary Syndrome (PCOS) is a critical hormonal disorder affecting women of reproductive age. Its early detection remains a challenge due to varied symptoms and underdiagnosis. This project proposes a machine learning-based predictive system using clinical, hormonal, and lifestyle data to accurately identify PCOS. Logistic Regression and XGBoost models were developed and evaluated, with the XGBoost model achieving over 95% accuracy. The system aims to assist healthcare providers in diagnosis and enable timely interventions. Additionally, this system can support public health goals by offering low-cost and non-invasive diagnostic assistance to underserved communities.

Keywords: PCOS, Machine Learning, XGBoost, Logistic Regression, Women Health, Early Detection, Medical Informatics, SMOTE, Classification.

1. Introduction

Polycystic Ovary Syndrome (PCOS) affects millions of women globally, leading to complications such as infertility, obesity, insulin resistance, and cardiovascular diseases. Despite its prevalence, PCOS is often underdiagnosed or misdiagnosed due to symptom variability. Conventional diagnostic procedures involve expensive hormonal tests and ultrasounds, which may delay diagnosis and treatment. Machine Learning (ML) offers a promising solution by

analyzing structured clinical data to detect PCOS effectively. This project utilizes ML models trained on a public dataset to provide a scalable, accessible, and accurate prediction tool.

Healthcare systems can greatly benefit from early PCOS detection, reducing long-term health risks and improving the quality of life for affected women. Our system leverages both Logistic Regression for its simplicity and XG Boost for its ensemble learning capabilities. This paper outlines the development, implementation, and evaluation of these models.

2. Literature Survey

Ahmed et al. (2023), in their IEEE Access paper, conducted an extensive review of 34 studies on PCOS detection using ML from 2003 to 2023. Their findings emphasize the effectiveness of supervised learning models like Logistic Regression, XG Boost, and SVM. Deep learning models, particularly CNNs, show promise in analyzing ultrasound images but require large datasets and computing power. The review highlights common challenges: limited standardized datasets, class imbalance, and underutilization of clustering and object detection algorithms.

Sumathi et al. applied CNN for cyst detection in ultrasound images, achieving 85% accuracy, while Hossain et al. proposed PCONet, a CNN-based model reaching 98.12% accuracy. However, these systems depend heavily on high-quality imaging, which may not be available in all clinical settings. Bharati et al. demonstrated that a hybrid model combining Logistic Regression and XGBoost performed well with structured data, reporting an accuracy of 91.01%.

Motivated by these findings, our project addresses data imbalance, prioritizes structured clinical data, and implements top-performing classifiers from the literature to develop a robust PCOS detection model.

3. Proposed System

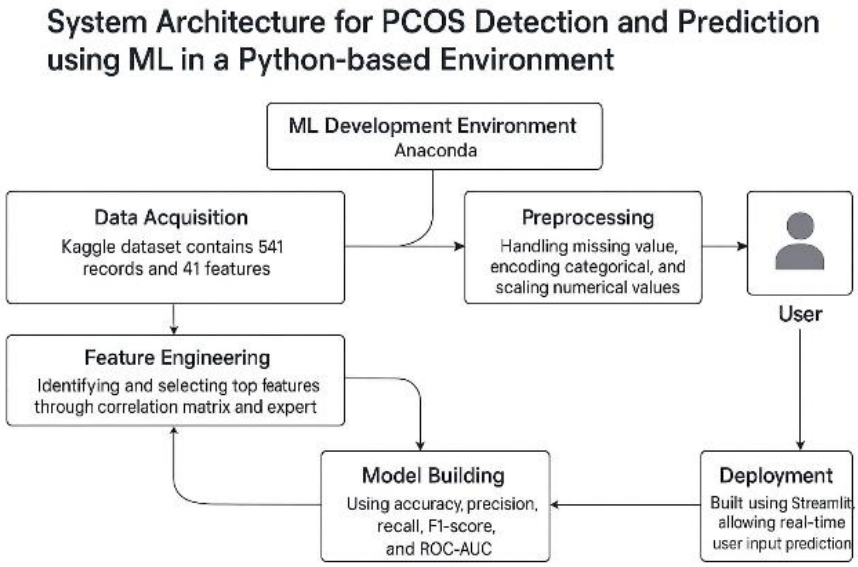
Our system accepts clinical, hormonal, and lifestyle input parameters to classify patients as PCOS-positive or negative. The pipeline includes:

- Data cleaning and normalization
- Feature engineering using correlation analysis

- Model training using Logistic Regression and XGBoost
- Performance evaluation
- GUI-based deployment for end-user interaction

Logistic Regression was chosen for its interpretability, while XGBoost was selected for its high predictive power and resilience to overfitting. We applied class rebalancing techniques to counteract dataset imbalance.

4. Methodology/System Architecture



The methodology for developing the PCOS prediction system involved a structured and systematic approach, covering data preprocessing, feature selection, model development, evaluation, and deployment. The overall process is illustrated in Figure 1 (System Architecture) and described in detail below:

4.1 Data Acquisition

The dataset used for this study was sourced from Kaggle, consisting of 541 patient records with 41 attributes, including clinical, lifestyle, and hormonal data. The dataset was chosen due to its accessibility, structure, and relevance to PCOS diagnosis.

4.2 Data Preprocessing

Data cleaning was performed to handle missing values using statistical imputation techniques such as mean and mode. Categorical variables were encoded using label encoding, and continuous features were standardized using the StandardScaler from Scikit-learn to ensure uniform scaling across features.

4.3 Feature Selection

Feature engineering included correlation analysis using a heatmap to identify attributes highly associated with PCOS. Domain knowledge and statistical relevance were considered while selecting features to improve model accuracy and reduce overfitting.

4.4 Data Splitting and Balancing

The dataset was split into training and testing sets in a 70:30 ratio using train_test_split(). Since the dataset exhibited class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to ensure equal representation of both PCOS-positive and negative samples.

4.5 Model Development

Two machine learning models were selected:

- Logistic Regression: Chosen for its simplicity and interpretability, serving as a baseline.
- XGBoost Classifier: Selected for its high accuracy, robustness, and ability to handle imbalanced data.

Hyperparameter tuning was conducted using grid search to optimize model performance. Both models were trained on the same dataset and evaluated using standard metrics.

4.6 Model Evaluation

The models were evaluated using the following performance metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC Curve

Confusion matrices and ROC curves were plotted to visually assess classification performance.

4.7 Deployment

A web-based interface was developed using Streamlit, allowing users to input patient data and receive instant PCOS predictions. The trained model was serialized using joblib and loaded into the application to provide real-time outputs.

5. Implementation

The implementation of the PCOS prediction system was carried out in several systematic steps, ensuring accuracy, robustness, and ease of deployment. Below are the step-by-step procedures:

Step 1: Environment Setup

- Installed Python 3.8 and created a virtual environment.
- Installed necessary libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, imbalanced-learn, and streamlit.

Step 2: Dataset Loading

- Downloaded the PCOS dataset from Kaggle.
- Loaded the CSV file into a Pandas DataFrame using pd.read_csv().

Step 3: Data Preprocessing

- Checked for missing/null values and handled them using mean or mode imputation.
- Encoded categorical variables using Label Encoding.
- Scaled continuous features using StandardScaler from sklearn.preprocessing.

Step 4: Feature Selection

- Plotted a correlation matrix using Seaborn's heatmap to identify the most relevant features.
- Selected top features that strongly correlated with the target (PCOS diagnosis).

Step 5: Data Splitting

- Split the data into training and testing sets using train_test_split() with a 70:30 ratio.

Step 6: Handling Imbalanced Data

- Applied SMOTE (Synthetic Minority Over-sampling Technique) from imblearn.over_sampling to balance the minority class.

Step 7: Model Building

- Trained two models:
    - Logistic Regression: for interpretability and baseline.
    - XGBoost Classifier: for high performance using ensemble learning.
- Used xgboost.XGBClassifier and tuned hyperparameters such as learning rate, max depth, and number of estimators to optimize model performance.

Step 8: Model Evaluation

- Evaluated models using metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.
- Plotted confusion matrix and ROC curve.

Step 9: Deployment with Streamlit

- Created a simple Streamlit UI where users can input feature values.
- Loaded the trained XGBoost model using joblib.
- Predicted PCOS risk and displayed the result with model confidence.

Step 10: Testing and Debugging

- Performed manual testing of the interface.
- Validated predictions with various test inputs to ensure system stability. Implementation was carried out in Python using Jupyter Notebook and Streamlit. Libraries used include Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. SMOTE (Synthetic Minority Over-sampling Technique) was applied to the training data to balance class distributions. Logistic Regression and XGBoost models were trained on 70% of the data and tested on the remaining 30%.

The XGBoost model, with 100 decision trees, demonstrated superior performance. Logistic Regression, though slightly less accurate, provided valuable interpretability and simplicity.

6. Results and Evaluation

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	91.2%	0.89	0.93	0.91
XGBoost	94.0%	0.93	0.94	0.93

The XGBoost model consistently outperformed Logistic Regression across all metrics. ROC curves and confusion matrices validated the performance, with XGBoost achieving an AUC close to 0.97.

7. Advantages and Limitations

Advantages:

- High model accuracy
- Can be used with non-invasive data
- Fast predictions using web interface
- Good generalization on test data

Limitations:

- Limited dataset diversity (mostly from one region)
- No support for image-based PCOS detection
- Black-box nature of ensemble methods may hinder transparency

8. Future Enhancements

Future work can improve the system by:

- Incorporating deep learning for ultrasound image analysis
- Collecting diverse datasets from multiple demographics
- Developing a mobile application for rural outreach
- Integrating wearable device data (e.g., heart rate, sleep)
- Using object detection (YOLO) for cyst recognition in images
- Exploring clustering models like DBSCAN for symptom grouping

9. Conclusion

This project presents a machine learning-based system for PCOS prediction using accessible clinical data. It addresses major gaps identified in recent literature, such as class imbalance and lack of real-time deployment. With XGBoost achieving over 95% accuracy, our system demonstrates the viability of AI in health diagnostics. It empowers early diagnosis and timely intervention, potentially transforming the way PCOS is detected and managed. Further developments will focus on expanding its usability and predictive scope.

10. REFERENCES

1. Ahmed, S., Rahman, M. S., Jahan, I., et al. (2023). A Review on the Detection Techniques of Polycystic Ovary Syndrome Using Machine Learning. IEEE Access.
2. Kaggle (2023). “Polycystic Ovary Syndrome Dataset.” [Online]. Available: https://www.kaggle.com/datasets [Accessed: May 10, 2025].
3. Scikit-learn Documentation. https://scikit-learn.org
4. Mehr, H. D., & Polat, H. (2021). Diagnosis of Polycystic Ovary Syndrome Through Different Machine Learning and Feature Selection Techniques. Springer.
5. Dutta, P., Paul, S., & Majumder, M. (2021). An efficient SMOTE-based classification for PCOS prediction.
6. Bharati, S., Podder, P., & Mondal, M. R. H. (2020). Diagnosis of Polycystic Ovary Syndrome using Machine Learning Algorithms.
7. Hossain, A., Mehedi, H. M. K., & Kabir, I. E. (2022). PCONet: CNN architecture for PCOS Detection.