International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 364-370, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE59
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

PCOS Detection Using Machine Learning

[1]G Dinesh Reddy, [2]Dr. M. Sireesh Kumar

[1]MCA student, Department of Computer Applications, Mohan Babu University, Tirupati, Andhra Pradesh, India
[2]Associate Professor, Department of Computer Applications, Mohan Babu University, Tirupati, Andhra Pradesh, India
E-mail: [1]dineeshreddy1029@gmail.com, [2]drmsk2102@gmail.com

***Abstract* - Reproductive-age women worldwide suffer from metabolic problems, hormonal imbalance, and irregular menstruation due to PCOS. Complex symptoms of PCOS often lead to misdiagnosis or underdiagnosis, causing suffering and increasing the risk of obesity, diabetes, and cardiovascular disease. Treating these symptoms requires early, correct diagnosis. The project tests machine learning techniques such as Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, K-Nearest Neighbors, Decision Tree Regressor, and Support Vector Machines employing Mean Squared Error to address diagnostic issues. Since it handled noisy and non-linear data better, the Random Forest Regressor was best. Django-based internet applications and predictive algorithm help clinicians identify PCOS risk. The approach instantly assesses risk using BMI, age, blood pressure, and lifestyle factors. This simple method lets clinicians identify high-risk patients for rapid intervention and personalized treatment. Accuracy, scalability, and usability tests validated the system's clinical value. Finally, our machine learning-based solution will improve early PCOS identification, clinical resource use, and global women's health.**

Keywords: PCOS, Machine Learning, Random Forest Regressor, Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Django, Prediction, Endocrine Disorder, Women's Health, Mean Squared Error, Real-time Application.

I. INTRODUCTION

A prevalent endocrine condition, PCOS, affects 6–10% of reproductive-age women [1]. PCOS' hormonal abnormalities, monthly irregularities, insulin resistance, and ovarian cysts affect women's mental and physical health. Obesity, hirsutism, infertility, and a higher risk of chronic diseases including diabetes and cardiovascular disease require early detection and treatment. PCOS is diverse, making clinical diagnosis difficult and sometimes erroneous. Patient history questionnaires, hormone testing, and ultrasonic imaging are expensive, invasive, and rare PCOS diagnosis procedures [2]. A faster, more accurate, and easier diagnostic tool is needed.

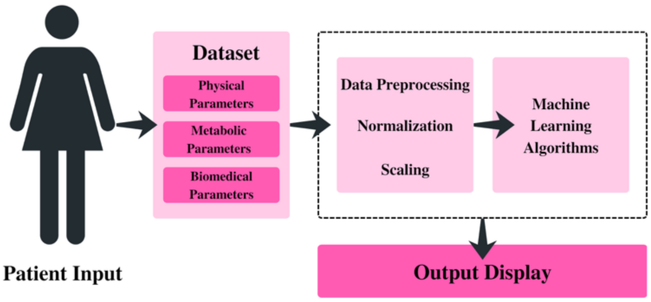


Figure 1: PCOS Detection

Modern artificial intelligence, such as machine learning (ML), can evaluate enormous volumes of data, uncover subtle patterns, and increase forecast accuracy [3]. These developments could greatly improve PCOS diagnosis. This study compares Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, K-Nearest Neighbors, Decision Tree Regressor, and SVM using Mean Squared Error (MSE) [4]. Random Forest Regressor was excellent for medical diagnostics since it could handle complex, non-linear data relationships and tolerate noise and outliers.

Health care practitioners can easily enter patient data and assess PCOS risk with the specified predictive risk model and Django web application. Age, BMI, blood pressure, diet, and exercise frequency are used to establish reliable early assessments without medical tests [5].

This technology aims to speed up diagnosis, improve early detection, remove invasive approaches, and optimize health care resource allocation. Doctors can better manage PCOS by providing precise and timely estimations, improving patient care and long-term health.

II. PROBLEM STATEMENT

Up to 10% of reproductive-age women globally have PCOS, which is harmful [6]. PCOS's intricacy and unpredictability lead to misdiagnosis and underdiagnosis. Symptoms including irregular menstruation, hormonal imbalance, insulin resistance, and ovarian cysts vary, making diagnosis and treatment difficult. The current diagnostic techniques, which include ultrasound imaging, patient history evaluations, and hormone testing, are costly, invasive, and not always accessible, especially in resource-constrained situations. Many affected individuals experience lengthy periods of untreated symptoms, which increases their risk of long-term health issues such as diabetes, obesity, cardiovascular disease, and infertility [7]. A reliable, accessible, and reasonably priced diagnostic tool that makes use of common clinical data is desperately needed to enable swift and early identification of PCOS. A machine learning predictive model with a user-friendly web interface based on machine learning is a viable technique to develop and use. This technique aims to significantly reduce diagnostic delays, streamline health care resources, and enhance patient outcomes using proactive and tailored therapies.

III. AIM & OBJECTIVES

For early PCOS identification and evaluation, the primary goal is to develop and implement a precise machine learning predictive model web application integrated within a user-friendly Django web application.

- Develop a predictive model for PCOS using multiple machine learning algorithms.
- Evaluate model performance using Mean Squared Error (MSE).
- Select the best-performing algorithm and integrate it into a Django-based web interface.
- Enhance diagnostic efficiency and accuracy using accessible clinical and lifestyle data.
- Validate model scalability, reliability, and practical utility through rigorous testing.

IV. RELATED WORKS

Much research has focused on PCOS prediction using Machine Learning (ML) techniques. But many research are performed and highlighted significance of machine learning (ML) in identifying trends that exisitng clinical diagnosis misses. Classification and regression models have been applied to PCOS datasets. This study demonstrated 85% PCOS diagnosis accuracy using SVM and Logistic Regression. Their research highlighted the significance of feature selection techniques, which significantly affected model performance.

Similar to this, a study [9] proposed an ensemble learning approach that merged Random Forest with Gradient Boosting to improve classification accuracy. Ensemble models outperform single classifiers because to their low overfitting and variance. Another study [10] looked into the application of deep learning for PCOS detection using Convolutional Neural Networks (CNN) to assess ultrasound images. With a 92% accuracy, the study shows the promise of image-based diagnostics combined with structured clinical data.

Engineering and feature selection have a significant role in ML model performance improvement. Singh and This work [11] investigated how regularity of the menstrual cycle, BMI, age, fasting insulin levels, and other elements affected model performance. Their study claims that combining physiological and behavioral elements improves predictive accuracy. In order to balance class distribution and improve classifier resilience, this study [12] also emphasized the significance of managing imbalanced datasets and applied SMote (Synthetic Minority Over-sampling Technique).

The integration of ML models into web-based applications has also been the subject of prior research. For instance, [13] developed a cloud-based PCOS diagnosis system using Flask, displaying good accessibility and usability for health care professionals. Their study revealed that combining a predictive machine learning model with web frameworks such as Django can enhance the applicability of predictive models in clinical contexts.

Notwithstanding these advancements, difficulties still exist, including model interpretability, dataset privacy concerns, and the need for large, diversified datasets for generalization. Several studies have concentrated on explainable artificial intelligence (XAI) techniques to improve the interpretability of ML predictions, making them more acceptable to clinicians. It will be imperative to solve these limitations through ongoing research and development in order to develop a reliable and trustworthy ML-driven PCOS screening system. This project includes a machine learning tool into a Django-based web application in order to provide a scalable and interpretable tool for PCOS risk assessment.

V. DATASET DESCRIPTION

This study used a dataset of 541 records with 44 variables, including PCOS-related demographic, clinical, and lifestyle factors. Important variables are age, weight, height, BMI, blood pressure, regularity of the menstrual cycle, hormone levels (FSH, LH, AMH, TSH), and lifestyle choices including nutrition and activity. The goal variable, determining if a patient has PCOS, is a binary indicator—PCOS: Yes/No. The dataset features numerical, category, and binary aspects, is well-structured, has few missing values. Preprocessing techniques, such as normalization, encoding, and handling missing values, are required to achieve the highest machine learning performance.

VI. METHODOLOGY

A. Data Preprocessing

The dataset underwent preprocessing to guarantee the accuracy and quality of the given data. Included were handling missing values, encoding category data, and standardizing numerical features. Mixed in the dataset called for feature scaling and transformation, constant and categorical elements ensure uniformity across all input variables.

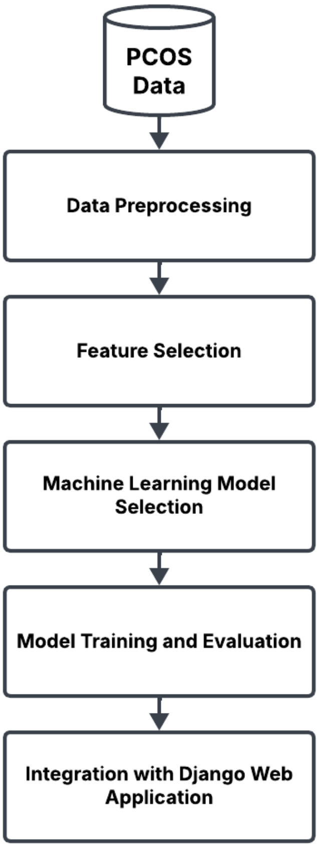

PCOS Data
Data Preprocessing
Feature Selection
Machine Learning Model Selection
Model Training and Evaluation
Integration with Django Web Application

Figure 2: Block Diagram

B. Feature Selection

Feature selection techniques were applied to improve model performance in order to uncover the most relevant PCOS diagnosis parameters. BMI, menstrual cycle length, FSH/LH ratio, and insulin resistance markers were eliminated using association analysis and recursive feature removal.

C. Machine Learning Model Selection

Some of the regression models investigated included Linear, Ridge, Lasso, Random Forest, k-nearest neighbors, Decision Tree, and SVM. Random Forest Regressor was selected as the best-performing model due to its resilience in managing non-linear correlations and feature importance. Using MSE, these models were evaluated.

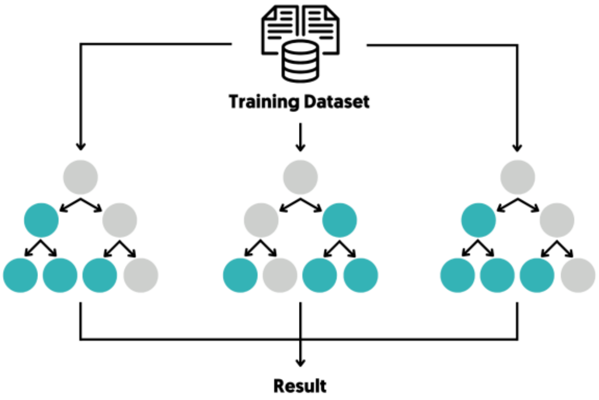


Figure 3: Random Forest

D. Model Training and Evaluation

Model Training (80%) and Model Testing (20%) dataset subsets for model validation and testing The Random Forest Regressor was trained on the training set and tested on the testing set using performance metrics including Mean Squared Error (MSE) and R-squared values. Cross-valuation prevented overfitting and supported model generalizability.

E. Integration with Django Web Application

The selected model was added to a Django-based Web Application to provide real-time PCOS risk assessment. The web interface allows users to enter patient health data, and the model then evaluates this data to generate an instantaneous PCOS likelihood prediction. The system provides ease of use for health care providers and can be implemented in clinical settings for efficient diagnosis and intervention.

This study uses machine learning to develop a scalable and effective PCOS diagnosis diagnostic tool. Incorporating Random Forest Regressor into a web application increases usability and accessibility, enabling quick PCOS diagnosis and treatment in a range of health care settings.

VII. RESULTS & DISCUSSION

Which model would be most fit was decided in part by mean squared error (MSE), which measures the average squared difference between the actual and projected values. A low MSE suggests a better-performance model. Many machines learning models' performance is displayed in the table below:

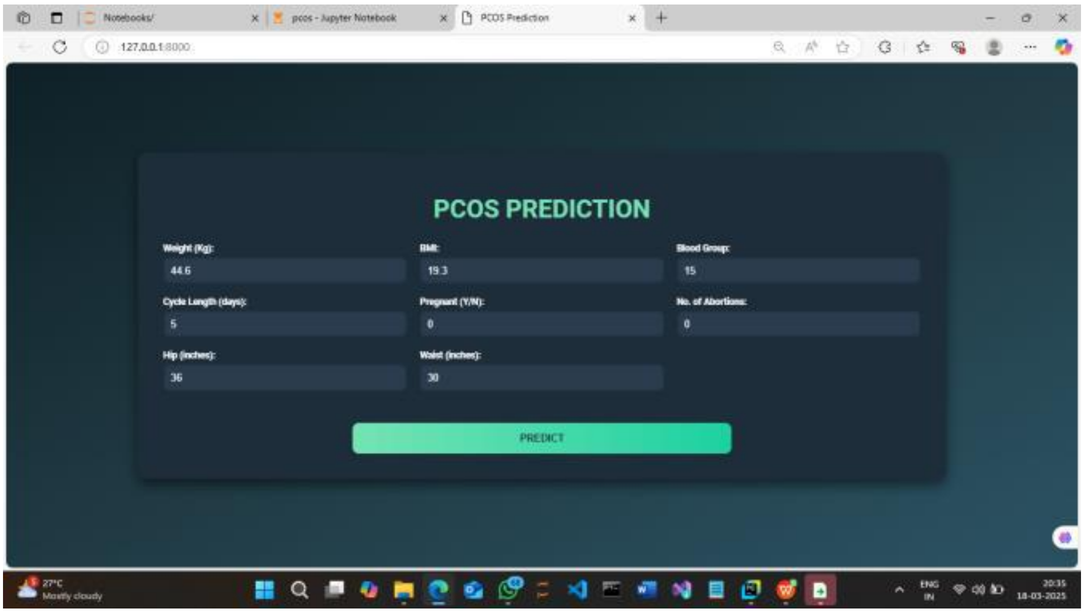
Model	Mean Squared Error (MSE)
Linear Regression	0.1004
Ridge Regression	0.1011
Lasso Regression	0.1619

Random Forest Regressor	0.0873
K-Nearest Neighbors	0.2333
Decision Tree Regressor	0.1667
Support Vector Machine	0.2237

Compared to other techniques, Random Forest Regressor has the lowest MSE (0.0873). Although their higher MSE values suggest they find it difficult to capture the non-linear trends in the dataset, Linear Regression and Ridge Regression also performed pretty nicely. Lasso Regression showed even greater inaccuracy due to its tendency to aggressively lower coefficients, making it less useful for this application. K-Nearest Neighbors and Support Vector Machine exposed poor generalization to unprocessed data using the best MSE values. Decision Tree Regressor lags Random Forest even if it is somewhat more sophisticated than KNN and SVM. The Random Forest Regressor was ultimately chosen because of its ability to control non-linearity, resistance to outliers, and high accuracy, making it the most suitable model for PCOS prediction.

This study uses machine learning to develop a scalable and effective PCOS diagnosis diagnostic tool. Incorporating Random Forest Regressor into a web application increases usability and accessibility, facilitating quick PCOS diagnosis and treatment in a variety of health care settings.

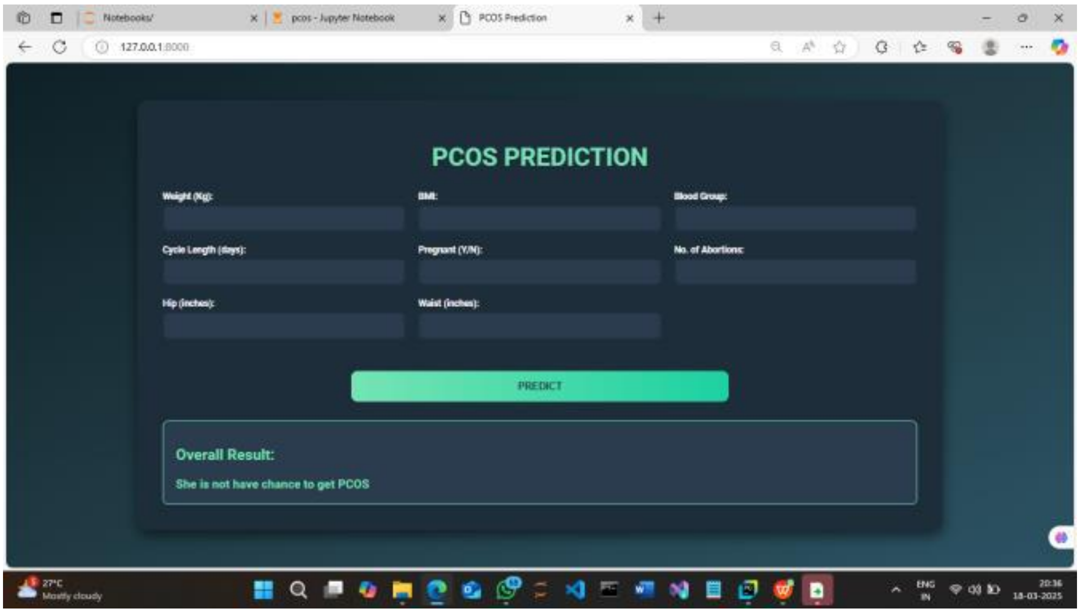
A. Deployment of web application



PCOS PREDICTION

Weight (Kg): 44.6
BMI: 19.3
Blood Group: 15
Cycle Length (days): 5
Pregnant (Y/N): 0
No. of Abortions: 0
Hip (inches): 36
Waist (inches): 30

PREDICT



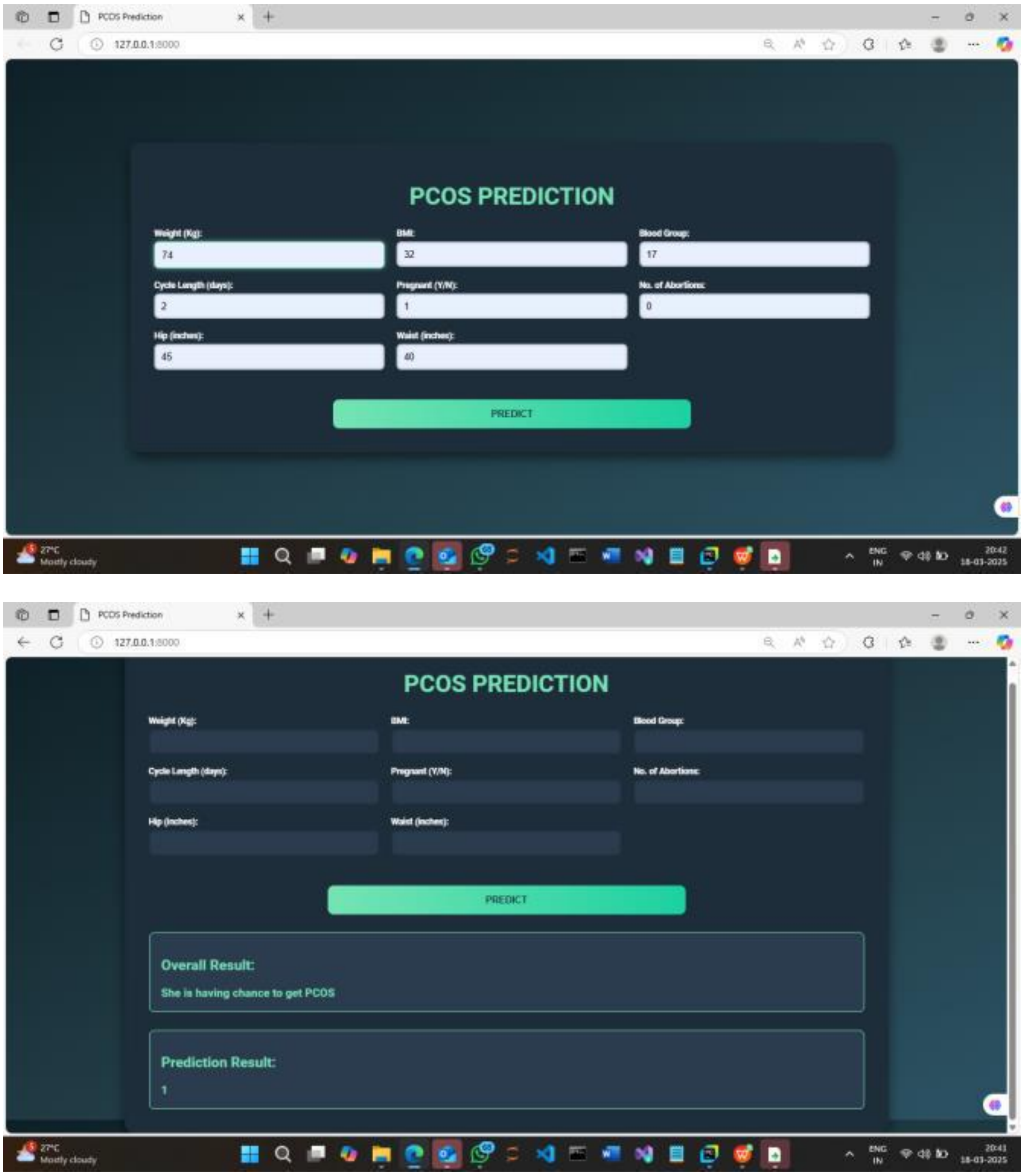
PCOS PREDICTION

Weight (Kg):
BMI:
Blood Group:
Cycle Length (days):
Pregnant (Y/N):
No. of Abortions:
Hip (inches):
Waist (inches):

PREDICT

Overall Result:

She is not have chance to get PCOS



PCOS PREDICTION

Weight (Kg): 74
BMI: 32
Blood Group: 17
Cycle Length (days): 2
Pregnant (Y/N): 1
No. of Abortions: 0
Hip (inches): 45
Waist (inches): 40

PREDICT

PCOS PREDICTION

Weight (Kg):
BMI:
Blood Group:
Cycle Length (days):
Pregnant (Y/N):
No. of Abortions:
Hip (inches):
Waist (inches):

PREDICT

Overall Result:
She is having chance to get PCOS

Prediction Result:
1

Figure 4: Predictions with UI

The PCOS Prediction Web Application is designed to aid in the early identification of Polycystic Ovary Syndrome and uses a Machine Learning Model incorporated into a Django framework. The system allows users to enter crucial health information such as weight, BMI, blood group, cycle duration, pregnancy status, abortion count, hip circumference, and waist size. Following the data being trained using the Random Forest Regressor model, the Predict button produces an instant prognosis. If a person has a low risk of developing PCOS, the system displays "She is not having a chance to get PCOS." Should the model show a strong probability, the message changes to "She is having a chance to get PCOS," with a 1 for the prediction result.

When no data is provided, the model defaults to a neutral state, displaying the empty form. The user provides values such as 44.6 kg weight, 19.3 BMI, 5-day cycle duration, 36-inch hip circumference, and 30-inch waist size in the second scenario. These inputs let the system determine there is no possibility of PCOS. The third scenario, on the other hand, has the user enter other parameters, such as 74 kg weight, 32 BMI, 2-day cycle duration, 45-inch hip circumference, and 40-inch waist size, which leads to the diagnosis of PCOS risk. The note "She is having a chance to get PCOS," validates this result.

Important clinical and behavioral factors let the web application effectively classify PCOS from non-PCOS cases. Machine learning application aids clinical decision making by enabling rapid identification and medical intervention. The Random Forest model was chosen for its non-linearity performance, feature priority, and noise tolerance. Its deployment in health care settings makes this web-based tool accessible to medical practitioners. With an automated, reliable, and scalable PCOS detection tool, this system demonstrates how machine learning may be applied to improve women's health.

VIII. CONCLUSION & FUTURE WORK

An early PCOS machine learning strategy based on a Random Forest Regressor model integrated inside a Django web application was successfully implemented in this application. The system provides a powerful and efficient diagnostic tool by combining clinical and lifestyle parameters to predict PCOS risk with high accuracy. Although the web interfaces provide accessibility for health care professionals, the model's ability to control complex linkages in data ensures predictions. By early PCOS detection and risk reduction, the system speeds medical therapy.

CNNs improve model performance in ultrasonic scan-based image-based diagnostics. By adding more demographic groups to the dataset, the model's generalizability will improve. The addition of Explainable AI (XAI) techniques can also enhance model transparency, making it easier for doctors to understand predictions. Clinical use will rise from better user interface design, mobile accessibility, and cloud deployment. At last, real-time information from wearable health sensors can enhance early diagnosis and best PCOS treatment.

REFERENCES

[1] S. B. M. R. Sokwala and R. Dodia, “Global approach to polycystic ovary syndrome in Africa,” in Elsevier eBooks, 2023, pp. 220–228. doi: 10.1016/b978-0-323-87932-3.00038-4. Available: https://doi.org/10.1016/b978-0-323-87932-3.00038-4.

[2] M. Thoma, J. Fledderjohann, C. Cox, and R. K. Adageba, “Biological and Social Aspects of Human Infertility: a Global perspective,” Oxford Research Encyclopedia of Global Public Health, Apr. 2019, doi: 10.1093/acrefore/9780190632366.013.184. Available: https://doi.org/10.1093/acrefore/9780190632366.013.184.

[3] I.H. Sarker, “Machine learning: algorithms, Real-World applications and research directions,” SN Computer Science, vol. 2, no. 3, Mar. 2021, doi: 10.1007/s42979-021-00592-x. Available: https://doi.org/10.1007/s42979-021-00592-x.

[4] M. Bansal, A. Raj, and A. Raj, “Comparative analysis of ML models for electricity price Forecasting,” in Lecture notes in networks and systems, 2024, pp. 551–578. doi: 10.1007/978-981-97-7710-5_42. Available: https://doi.org/10.1007/978-981-97-7710-5_42.

[5] M. Almutairi, A. A. Almutairi, and A. M. Alodhialah, “The Influence of Lifestyle Modifications on Cardiovascular Outcomes in Older Adults: Findings from a Cross-Sectional Study,” Life, vol. 15, no. 1, p. 87, Jan. 2025, doi: 10.3390/life15010087. Available: https://doi.org/10.3390/life15010087.

[6] D. Vine, M. Ghosh, T. Wang, and J. Bakal, “Increased prevalence of adverse health outcomes across the lifespan in those affected by polycystic ovary syndrome: a Canadian population cohort,” CJC Open, vol. 6, no. 2, pp. 314–326, Dec. 2023, doi: 10.1016/j.cjco.2023.12.010. Available: https://doi.org/10.1016/j.cjco.2023.12.010.

[7] C. Arslanian-Engoren, R. Gary, C. Irwin, and W. Zhang, “Chronic and other conditions that increase CVD risk,” in Springer eBooks, 2024, pp. 181–227. doi: 10.1007/978-3-031-53705-9_7. Available: https://doi.org/10.1007/978-3-031-53705-9_7.

[8] V. Gupta and P. V. Suresh, “A comprehensive review of predicting Lifestyle-Based Disease specifically PCOS among women using data mining and machine learning approaches,” Lecture Notes in Networks and Systems, pp. 419–433, Jan. 2024, doi: 10.1007/978-981-97-2089-7_37. Available: https://doi.org/10.1007/978-981-97-2089-7_37.

[9] T. Kavzoglu and A. Teke, “Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBOOST) and natural gradient boosting (NGBOOST),” Arabian Journal for Science and Engineering, vol. 47, no. 6, pp. 7367–7385, Jan. 2022, doi: 10.1007/s13369-022-06560-8. Available: https://doi.org/10.1007/s13369-022-06560-8.

[10] A.Alamoudi et al., “A deep learning fusion approach to diagnosis the polycystic ovary syndrome (PCOS),” Applied Computational Intelligence and Soft Computing, vol. 2023, pp. 1–15, Feb. 2023, doi: 10.1155/2023/9686697. Available: https://doi.org/10.1155/2023/9686697.

[11] K. Hussein and M. Karami, “Association between insulin resistance and abnormal menstrual cycle in Saudi females with polycystic ovary syndrome,” Saudi Pharmaceutical Journal, vol. 31, no. 6, pp. 1104–1108, Apr. 2023, doi: 10.1016/j.jsps.2023.03.021. Available: https://doi.org/10.1016/j.jsps.2023.03.021.

[12] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, “Effective Class-Imbalance

learning based on SMOTE and convolutional neural networks," Applied Sciences, vol. 13, no. 6, p. 4006, Mar. 2023, doi: 10.3390/app13064006. Available: https://doi.org/10.3390/app13064006.

[13] C. C. Yang, "Explainable artificial intelligence for predictive modeling in healthcare," Journal of Healthcare Informatics Research, vol. 6, no. 2, pp. 228–239, Feb. 2022, doi: 10.1007/s41666-022-00114-1. Available: https://doi.org/10.1007/s41666-022-00114-1.

Citation of this Article:

G Dinesh Reddy, & Dr. M. Sireesh Kumar. (2025). PCOS Detection Using Machine Learning. In proceeding of International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25), published by *IRJIET*, Volume 9, Special Issue of INSPIRE'25, pp 364-370. Article DOI https://doi.org/10.47001/IRJIET/2025.INSPIRE59

****