OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21610366

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Articles

Volume-06|Issue-04|2026

Leveraging Machine Learning to Improve PCOS Detection through Ovarian Analysis

Dr. B P Pradeep Kumar[1] and Manjunath K M[2]

[1]Professor, Computer Science and Design, Atria Institute of Technology, Bengaluru, Karnataka, India.
[2]Assistant Professor, Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, Karnataka, India.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Kumar, B. P. P. & Manjunath, K. M. (2026) Leveraging Machine Learning to Improve PCOS Detection through Ovarian Analysis. *Indiana Journal of Multidisciplinary Research, 6*(4), 404-408.

Abstract: A complex endocrine condition, PCOS affects up to 10% of women who are of reproductive age. Infertility and metabolic problems can be avoided with early diagnosis. For PCOS prediction, this study suggests a machine learning (ML)-based diagnostic approach that makes use of clinical, biochemical, and lifestyle variables. Using the Kaggle PCOS dataset, which included 541 samples with 42 characteristics, models including Random Forest, Support Vector Machine (SVM), and Neural Networks were trained. Following 10-fold cross-validation and Randomized Search hyperparameter adjustment, the Random Forest classifier produced an accuracy of 87.73%, a macro F1-score of 0.88, and an AUC-ROC of 0.93. Weight gain, hair growth, and follicle count were found to be the most significant predictors by SHAP analysis. The clinical promise of ML-based screening for early, non-invasive, and interpretable PCOS identification is demonstrated by these findings.

Keywords: Polycystic ovaries, Biochemical, Endocrine conditions predictive, Support Machine Vector, non-invasive.

INTRODUCTION

An estimated 6–10% of women worldwide who are of reproductive age suffer from PCOS, a multifactorial endocrine condition. PCOS has serious effects on metabolic, psychological, and reproductive health. It is characterized by a variety of symptoms, such as polycystic ovarian morphology, hyperandrogenism, and oligo- or anovulation. Infertility as diabetes type 2, heart disease, and endometrial cancer are among the problems that can be prevented with early diagnosis and intervention-[21]. Conventional methods of diagnosing PCOS, such the Rotterdam criteria24, depend on assessments that are clinical, biochemical, and ultrasound-based. Nevertheless, these approaches are frequently arbitrary, time-consuming, and constrained by interpretational variability[22]. Recent developments in artificial intelligence (AI) and medical imaging have opened up new possibilities for increasing the precision and effectiveness of PCOS diagnosis[23].

A subset of artificial intelligence called machine learning (ML) has shown great promise in the healthcare industry for decision assistance and predictive analytics systems [24]. Large datasets with medically biochemical, and imaging information can be analysed by ML algorithms to find patterns and relationships that may not be visible using conventional techniques in connection with PCOS detection[25]. Based on patient-specific data, methods such neural network designs, supervised learning, and feature selection have demonstrated promise in PCOS prediction[26]. By combining medically biochemical, and imaging data, our goal in this project is to create a machine learning-based diagnosis paradigm for PCOS. To improve diagnostic precision and offer tailored insights, the suggested method makes use of feature selection and sophisticated machine learning algorithms. This strategy emphasizes the role of AI in improving women's healthcare while simultaneously attempting to solve the shortcomings of conventional diagnostic techniques [27].

Inter-observer variability and varied symptom presentation make PCOS detection unreliable even with advances in diagnostic imaging and endocrinology. There is a chance to automate feature extraction and standardize diagnosis by combining machine learning with biochemical and ovarian data. In order to increase diagnostic confidence and aid in therapeutic decision-making, this study focuses on integrating ovarian follicle measures, hormonal factors, and lifestyle characteristics into an interpretable machine learning pipeline.

***Corresponding Author:** Dr. B P Pradeep Kumar

MATERIALS AND METHODS

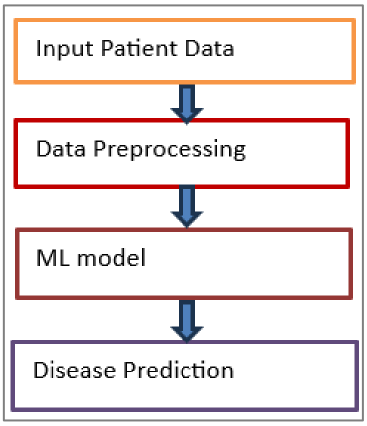

Fig. 1: Block diagram for multiple disease prediction using ML

The methodology shown in figure 1 illustrates Multiple Disease Prediction System outlines the systematic approach adopted to design, develop, and implement the solution using machine learning techniques.The goal of this study's methodology is to create a precise and effective diagnosis model for polycystic ovarian syndrome (PCOS) by utilizing machine learning (ML) techniques. To guarantee a thorough dataset that captures the complex character of PCOS, the procedure starts with data gathering from a variety of sources, such as clinical records, biochemical profiles, hormone levels, and ultrasound imaging data. Data preprocessing, which includes handling missing values, standardizing continuous variables, and resolving class imbalances using strategies like the Synthetic Minority Over-sampling Technique (SMOTE), is essential given the disorder's variability.

Data Collection:
The dataset for this study was sourced from the Kaggle platform (https://www.kaggle.com/code/karnikakapoor/pcos-diagnosis) under the title "PCOS Diagnosis," curated by Karnika Kapoor.It includes patient demographic and clinical data that has been anonymised and is intended especially for the investigation and diagnosis of PCOS. A vast array of characteristics, including age, BMI, irregular menstrual cycles, hormone profiles, ultrasound findings, and family history of PCOS, are included in the dataset, which comprises 541 samples with 42 variables. This thorough data makes it possible to compare studies between PCOS patients and healthy controls as well as to thoroughly examine the elements that contribute to the condition.

To guarantee integrity and dependability for analysis, the data underwent extensive preparation. In order to reduce potential biases, outliers were evaluated and missing values were imputed using statistical techniques. To further improve interoperability with machine learning techniques, all numerical variables were standardized. Since all records have been de-identified and anonymised to preserve patient privacy, the dataset complies with ethical standards. This dataset provides a strong basis for investigating PCOS diagnosis and prediction methods.

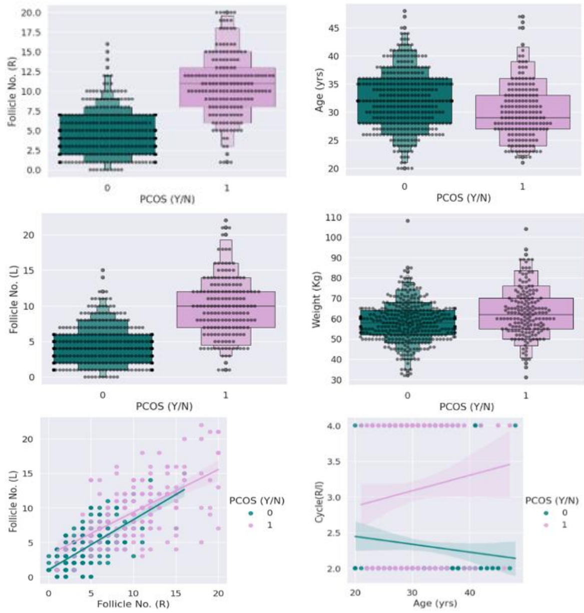

Fig. 2: Data visualization of the features

Figure 2 displays six data visualization charts that highlight important characteristics associated with PCOS: left and right ovarian follicle counts indicate polycystic morphology; hair growth indicates hyperandrogenism; weight gain indicates metabolic imbalance; lifestyle patterns, such as fast-food consumption, indicate indirect influence; and age distribution shows higher prevalence in reproductive age groups. When combined, these graphs offer a comprehensive summary of the behavioral, physiological, hormonal, and demographic variables most important for PCOS prediction.

In order to guarantee the precision and efficacy of machine learning algorithms for PCOS detection, data preparation is essential. The first step in preprocessing is data cleaning, which involves addressing missing values in the dataset using imputation approaches such the most frequent subcategory for categorical variables and the median, mean, or mode impute for numerical features. Because they can distort model findings, outliers are also found and dealt with. In order to make categorical variables useful with machine learning algorithms, the next step is feature encoding, which converts Feature selection was carried out by combining statistical techniques with the Random Forest algorithm's feature priority attribute in order to enhance model performance and decrease dimensionality. To ensure that only the most significant markers of PCOS could be included in the final model, the most pertinent characteristics were ranked using Recursive Feature Elimination (RFE). This phase improved the interpretability of the model and

***Corresponding Author:** Dr. B P Pradeep Kumar

prevented overfitting.The Random Forest classification system was selected as the main model for identifying Polycystic Ovaries Syndrome in this study because of its resilience, adaptability. Randomized Search was used to optimize the Random Forest classifier's hyperparameters in the suggested machine learning approach for identifying Polycystic Ovary Syndrome (PCOS). A crucial step in enhancing machine learning models' performance and capacity for generalization is hyperparameter tuning.

By randomly selecting combinations from a predetermined search area, Randomized Search offers an effective method for investigating a broad variety of hyperparameter values. Because Randomized Search selects a defined number of parameter values rather than testing every conceivable combination like Grid Search does, it is less computationally prohibitive while encompassing a large portion of the hyper parameter space.

On the basis of the validation data set, the effectiveness of the machine learning technique for identifying PCOS was assessed. With a validation accuracy of 87.73%, the model proved to be highly effective in accurately classifying PCOS cases. This high accuracy suggests that the Random Forest classifier was able to efficiently learn from the clinical and biochemical aspects of the dataset, providing accurate predictions, after being refined via Randomized Search. Furthermore, the loss of validity was 0.3072, demonstrating the model's capacity to reduce mistakes throughout testing and training—a critical component in guaranteeing the model's generalizability and resilience.

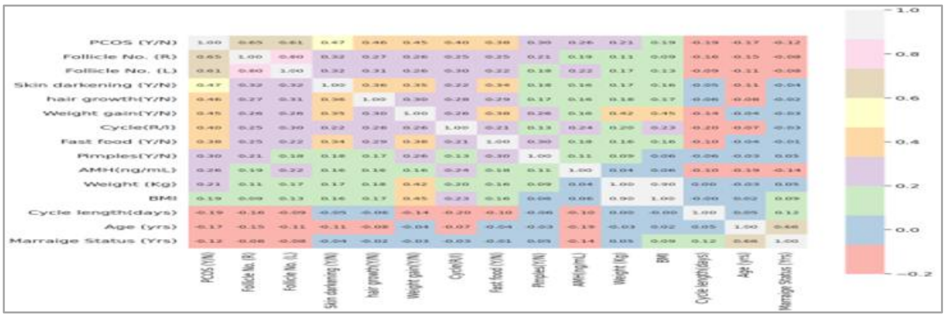

Fig. 3: Correlation matrix of PCOS disease

The relationship between several clinical and lifestyle characteristics linked to Polycystic Ovary Syndrome (PCOS) is depicted in the heatmap in Figure 3. Warmer hues indicate positive correlations, whereas cooler hues indicate negative correlations. The number of left and right ovarian follicles is strongly correlated, indicating that these characteristics are interdependent in ovarian morphology. Weight increase, hair growth, and follicle count all exhibit moderately favorable relationships with PCOS occurrence, indicating their significance in the syndrome's expression. Furthermore, there are weak associations between lifestyle factors like eating fast food and darkening of the skin, suggesting that they may have an indirect effect. In contrast, there are weak or negative associations between PCOS and characteristics including age, BMI, and marital status, suggesting that these variables have less predictive power in this sample.

RESULTS

Polycystic Ovary Syndrome (PCOS) was examined using precision, recall, and The F1 score metrics, with findings demonstrating strong model efficacy. The model's great capacity to accurately identify non-PCOS patients was demonstrated by its precision of 0.89, recall for 0.95, and F1-score of 0.92 for category 0 (non-PCOS cases). The precision, recall, and F1-score for class 1 (symptoms of PCOS cases) were 0.87, 0.75, and 0.81, respectively. The classifier did a good job of detecting non-PCOS instances, but it did a little worse on PCOS cases, suggesting that there may be space for improvement for PCOS diagnosis. The model's macro average reliability, recollection, and F1-score were 0.88, 0.85, and 0.86, respectively, indicating an overall accuracy of 88%.

A confusion matrix was used to assess the effectiveness of the suggested machine learning algorithm for PCOS detection, producing metrics for precision, recall, and F1-score for the PCOS-positive and PCOS-negative classes. The model showed a solid recognition of non-PCOS cases for class 0 (PCOS-negative), with an F1-score of 0.92, a high precision of 0.89, and a great recall of 0.95. The model's accuracy and recall for class 1 (PCOS-positive) were 0.87 and 0.75, respectively, yielding an F1-score of 0.81. This figure 4 demonstrates how the model's sensitivity for identifying affirmative situations is marginally lower than its precision.

Fig 4: Correlation matrix of PCOS disease

***Corresponding Author:** Dr. B P Pradeep Kumar

CONCLUSION

Given the complexity of Polycystic Ovary Syndrome (PCOS), which has many symptoms and major health implications, a timely and accurate diagnosis is essential. By examining intricate patterns in clinical, biochemical, and imaging data, machine learning (ML) has become a successful method for identifying PCOS. By using sophisticated algorithms like decision trees, support vector machines, and neural networks, machine learning (ML) increases diagnostic accuracy above conventional techniques while also handling non-linear correlations and multidimensional data with ease. Furthermore, ML provides early PCOS detection, reducing the risk of complications and enabling earlier treatments. Additionally, by shedding light on important prognostic indicators like hormone levels and body mass index, these models provide a deeper understanding of the illness and the creation of individualized diagnostic and therapeutic approaches.

REFERENCES

1. Augustine, R. (2025). *Efficient deep learning architectures for brain tumor diagnosis: A comparative study of Xception, MobileNet, MLUNet and DeepLabS* (SSRN Working Paper No. 5236459). SSRN. https://doi.org/10.2139/ssrn.5236459
2. Augustine, R. (2025). *Machine learning for personalized juvenile diabetes risk stratification: An in-depth comparative study* (SSRN Working Paper No. 5177119). SSRN. https://ssrn.com/abstract=5177119
3. Rangaiah, P. K. B., Huss, F., et al. (2025). Precision diagnosis of burn injuries using imaging and predictive modeling for clinical applications. *Scientific Reports, 15*, Article 7604. https://doi.org/10.1038/s41598-025-92096-4
4. Shanthamallappa, M., & Pradeep Kumar, B. P. (2025). Enhanced perceptual wavelet packet features for spontaneous Kannada sentence recognition under uncontrolled conditions. *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-024-10156-y
5. Naresh, E., Ashwitha, A., Reddy, K. T., & Srinidhi, N. N. (2025). The burn grafting image reclamation redefined with the peak-valley approach. *Critical Reviews in Biomedical Engineering, 53*(2).
6. Kumar, B. P. P., et al. (2022). Advanced electric vehicle charging infrastructure using Internet of Things. In *2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4)* (pp. 1–6). IEEE. https://doi.org/10.1109/C2I456876.2022.10051342
7. Naresh, E., Murthy, S. V. N., Pareek, P. K., Reddy, K. T., Shiva Darshan, S. L., et al. (2024). DevOps life cycle implementation on real life scenarios. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1269–1271). IEEE. https://doi.org/10.1109/ICKECS61492.2024.10617200
8. V. C., R., S. A., D., B., & V., S. (2023). Speech-to-text transfiguration in language numerals for perpetual deaf patients. In *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICEEICT56924.2023.10157113
9. Surendiran, J., Reenadevi, R., Vidhya, R. G., Sivasankari, S. S., & Balaji, N. (2022). IoT-based advanced electric vehicle charging infrastructure. In *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)* (pp. 1–6). IEEE. https://doi.org/10.1109/CCIP57447.2022.10058649
10. Naresh, E., Raghavendra, C. K., et al. (2025). Comprehensive brain tumour concealment utilizing peak valley filtering and DeepLab segmentation. *Scientific Reports, 15*, Article 34780. https://doi.org/10.1038/s41598-025-18574-x
11. Sheela, S., Jyothi, S., & Kumar, A. (2025). A hybrid approach for vehicular spectrum allocation using artificial neural networks with autoencoders and CSI feedback. *SN Computer Science, 6*, Article 877. https://doi.org/10.1007/s42979-025-04413-3
12. 2
13. Kousar, F., Namratha, R. N., & Yogesh, D. (2025). Leveraging AI for advancing health and behavioral optimization. In *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICKECS65700.2025.11035259
14. Hari Gokul, S. S., Gow, A., & Adarsh, L. (2025). Remote monitoring for water level of bridges and flood zones. In *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICKECS65700.2025.11035864
15. John, A., Harshitha, K., & Pooja, D. (2025). Leveraging histopathological images for detection and diagnosis of colonic carcinoma. In *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICKECS65700.2025.11035505
16. Tabassum, M., Neha, R., & Khanum, A. (2025). *ChargeSmart: Effortless EV charging with real-time slot booking*. In *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICKECS65700.2025.11035724
17. Purushotham, R., Mahato, R., & Kumar, S. M. V. (2025). AI-powered plagiarism detection: A web-based system for text integrity. In *2025 International*

***Corresponding Author:** Dr. B P Pradeep Kumar

Conference on Knowledge Engineering and Communication Systems (ICKECS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICKECS65700.2025.11036044
18. Afnar, R., Raheem, R., Ranjanyou, N., Ravikumar, J., & Veena, S. (2025). Hand motions based virtual AI mouse. In *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICKECS65700.2025.11035082
19. Harshvardhan, M., Devineni, M., & Raju, P. A. (2025). Framework for tuberculosis detection. In *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICKECS65700.2025.11035324
20. Ravikumar, J., Shankar, B. B., & Manjunath Kamath, K. (2025). Optimized anti-interference dynamic integral neural network approach for dementia prediction in health care. *Knowledge-Based Systems, 321*, Article 113723. https://doi.org/10.1016/j.knosys.2025.113723
21. Rangaiah, P. K. B., & Augustine. (2025). From microwave measurement to application: Enhancement of fat-intrabody communication by advanced computational techniques. In *2025 16th German Microwave Conference (GeMiC)* (pp. 490–493). IEEE. https://doi.org/10.23919/GeMiC64734.2025.10979071
22. Pradeep Kumar, B. P. (2025). A comparison of CNN, RNN, and FNN algorithms to investigate effective diabetes prediction. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-025-10333-5
23. Naresh, E., Hemavathi, P., Padmavathi, S., Srinidhi, N. N., Pradeep Kumar, B. P., Karthik, V., & Mallik, S. (2024). Autonomous garbage accumulation robot using Internet of Things. *Journal of Machine and Computing, 4*(2), 431–440.
24. R., L., R., S. A., Ibrahim, M. M., & V., S. (2023). Hybrid threshold speech enhancement scheme using TEO and wavelet coefficients. In *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICEEICT56924.2023.10156921
25. Kumar, B. P. (2018). Analysis of ASL recognition system for aligned RGB-D image. *Journal of Advanced Research in Dynamical and Control Systems*.
26. Mallanna, S. D. (2018). High gain perfect matched inset-fed rectangular microstrip patch antenna for 2.4 GHz frequency. *Journal of Advanced Research in Dynamical and Control Systems, 10*(13, Special Issue), 46–52.
27. Pramod, K. B., Kumaraswamy, H. V., Prathap, C., & Swamy, M. (2013). Design and analysis of UHF BJT feedback oscillator using linear and non-linear simulation. In *2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA)* (pp. 1–6). IEEE. https://doi.org/10.1109/C2SPCA.2013.6749386
28. Ahmed, M. (2025). IoT-based battery management system in the electric vehicle charging station. In T. A. Nguyen (Ed.), *Digital twin, blockchain, and sensor networks in the healthy and mobile city* (pp. 431–451). Elsevier. https://doi.org/10.1016/B978-0-443-34174-8.00022-3
29. Somyaji, A. D., Surakasi, L., Puranik, A., & Kannan, D. (2025). Melody mixer: Novel music generation by mixing styles with variational autoencoders. In *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)* (pp. 1–5). IEEE. https://doi.org/10.1109/COMP-SIF65618.2025.10969893
30. D. J., B., G. G., & M. N. (2025). Fake job post detection using machine learning. In *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)* (pp. 1–6). IEEE. https://doi.org/10.1109/COMP-SIF65618.2025.10969867
31. Chand, S. R., Ahmed, U., Ismail, T. M., Chand, S. B., & P. C. (2025). Email validator dashboard: A comprehensive tool for accurate email validation. In *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)* (pp. 1–8). IEEE. https://doi.org/10.1109/COMP-SIF65618.2025.10969888
32. Kumar, R. J., Biswas, A., Hafeeza, B. B., & Dhaanya, D. (2025). Emotion classification in text using machine learning and LSTM. In *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)* (pp. 1–7). IEEE. https://doi.org/10.1109/COMP-SIF65618.2025.10969881
33. Khan, R. M., Kabir, S., & Raj, Y. (2025). Driver drowsiness detection system. In *2025 3rd International Conference on Smart Systems for Applications in Electrical Sciences (ICSSES)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICSSES64899.2025.11010064
34. Reddy, L. G., Chethana, M., & Likhita, D. (2025). Improving DDoS attack identification and mitigation in SDN through the use of ensemble online machine learning framework. In *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICKECS65700.2025.11035808

***Corresponding Author:** Dr. B P Pradeep Kumar