ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug048

Development of a Random Forest-Based Predictive Model for Polycystic Ovary Syndrome (PCOS) Using SHAP for Explanability

Chidimma Nweke[1*]; Prema Kirubakaran[2]; Ridwan Kolapo[3]

[1;2;3]Department of Information Technology, Faculty of Computing, Nile University of Nigeria, Abuja 900108, Nigeria

Corresponding Author: Chidimma Nweke[1*]

Publication Date: 2026/08/08

Abstract: Polycystic Ovary Syndrome (PCOS) is a lead major disorder and primary cause of infertility in women of reproductive age, affecting about 13% of this population globally with over 70% of cases remaining undiagnosed. Early diagnosis is yet challenging, particularly in low-resource settings where ultrasound imaging is inaccessible. This study focuses on leveraging Random Forest (RF) model for PCOS prediction using only clinical and biochemical data, enhanced with Shapley Additive Explanations (SHAP) for model interpretability. A publicly available Kaggle PCOS dataset from 541 women (177 PCOS-positive, 364 negative) across 10 hospitals in Kerala, India, was utilised. A leakage-free preprocessing pipeline applied median and mode imputation before data splitting. SelectKBest (ANOVA F-test) was utilized. Hyperparameter optimization employed Randomized Search CV (75 iterations) with Repeated StratifiedKFold cross-validation (5 splits x 2 repeats). A custom probability threshold of 0.46 balanced sensitivity and specificity. The optimised model achieved 92% accuracy, 91% precision, 83% recall, 87% F1-score, 95.8% specificity, and ROC-AUC of 0.954. SHAP analysis identified weight gain (0.086), skin darkening (0.078), and AMH levels (0.067) as the top three predictors, aligned with the Rotterdam Criteria. The model outperforms comparable non-imaging approaches and demonstrates that accurate, explainable PCOS diagnosis is achievable without imaging data, making it suitable for resource-limited healthcare settings.

Keywords: *Polycystic Ovary Syndrome, Random Forest, Shapley Additive Explanations, Machine Learning, Explainable Artificial Intelligence, Clinical Diagnosis.*

How to Cite: Chidimma Nweke; Prema Kirubakaran; Ridwan Kolapo (2026) Development of a Random Forest-Based Predictive Model for Polycystic Ovary Syndrome (PCOS) Using SHAP for Explanability. *International Journal of Innovative Science and Research Technology*, 11(8), 88-96. https://doi.org/10.38124/ijisrt/26aug048

I. INTRODUCTION

Polycystic Ovary Syndrome PCOS is a complex hormonal and metabolic disorder affecting approximately 10-13% of women of reproductive age worldwide, with up to 70% of cases remaining undiagnosed [1][2]. It is characterized by hormonal imbalances, hyperandrogenism, irregular menstrual cycles, and anovulation, and is a leading cause of female infertility [3]. PCOS is also associated with long-term comorbidities including insulin resistance, obesity, type 2 diabetes, and cardiovascular disease [4].

Despite its prevalence, timely diagnosis remains elusive. The Rotterdam Criteria, requiring at least two of three indicators (hyperandrogenism, ovulatory dysfunction, and polycystic ovarian morphology), are consider to be the standard diagnostic framework [5]. Conventional diagnostic pathways depend on transvaginal ultrasound, hormonal assays, and clinical examination, which are resource-intensive, operator-dependent, and inaccessible in many low-resource healthcare environments [6].

In recent years, the Machine learning model has demonstrated significant potential in improving diagnostic accuracy across medical domains [7]. Several of these models for PCOS detection have yielded high accuracy; however, many rely on imaging-derived features [8][9], limiting applicability in settings lacking ultrasonography infrastructure. Furthermore, the black-box nature of many high-performing models restricts clinical adoption, as clinicians lack insight into predictive logic [10]. Moreso, the Explainable Artificial Intelligence (XAI), particularly Shapley Additive Explanations (SHAP), has emerged as a major tool for enhancing transparency and interpretability in

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug048

clinical ML models [11]. [12] demonstrated the feasibility of combining RF with SHAP for PCOS prediction using metabolomic follicular fluid data, achieving 86% accuracy. However, the reliance on invasive specimen collection and a small sample (n=72) limits generalizability and clinical utility.

This study addresses these gaps by: (i) designing a refined RF model using an optimized hyperparameter strategy with only non-imaging clinical and biochemical features; (ii) implementing the model in a Python-based ML framework; (iii) incorporating SHAP for model explainability; and (iv) evaluating model performance using metrics like accuracy, precision, recall, F1-score, specificity, and roc-auc. This study targets the dual challenge of accessibility and interpretability in PCOS diagnostics, supporting early detection in resource-limited settings.

II. RELATED WORKS

[13] developed an intelligence system to classify PCOS using k-means coupled with Least Squares SVM (LS-SVM), achieving an accuracy of 99%. Although these models are statistically strong, they face challenges in clinical settings where justification for treatment is needed due to the lack of explainability methods.

[14] used a sizable dataset of 30,601 electronic health records (EHRs) on women's clinical, biochemical, and transabdominal or transvaginal ultrasound tests. Their techniques, which included RFs, Gradient-Boosted Trees (GBTs), SVMs, and LR, produced AUCs ranging from 77.4% to 85%, with GBT being the best performer. However, despite the large dataset, the findings cannot be generalized due to a lack of external validation. Additionally, the absence of a practical deployment framework and its restriction to a geographical region limit its application in clinical settings and create biases.

Systematic reviews by [15] and [6] report the importance and significance of ML models in PCOS diagnosis. These studies highlighted the high diagnostic efficiency and precision of these models. However, they also noted the lack of transparency in the models. Another limitation observed was that most of the studies reviewed relied on small localized datasets, which limits the generalizability of the models to the wider global population.

Some studies focused on improving feature selection to determine the most relevant predictors for PCOS. The review by [16] reported that high-performing models used in PCOS diagnosis used features derived from transvaginal ultrasound, such as follicle count or ovarian volume. RFs and SVMs achieved highly consistent performance in clinical and biochemical datasets. However, it also reported that only a quarter of the studies reviewed applied XAI methods, limiting transparency and trust.

Finally, [12] Yasar, (2025) conducted a study to predict PCOS using metabolomic data from follicular fluid and the Random Forest algorithm and SHAP for explainability. The RF model attained an accuracy of 0.86. SHAP analysis identified L-Histidine, L-Glutamine, and L-Tyrosine as the most influential metabolites associated with PCOS. This study illustrates the potential of integrating ML with XAI to accurately predict PCOS without imaging. The approach could improve diagnostic accuracy and accessibility, especially in low-resource settings.

III. MATERIALS AND METHODS

- *Study Design and Framework*

This study employed a quantitative experimental design using supervised ML. The overall research framework, illustrated in Figure 1, encompasses nine sequential stages: data collection, preprocessing, feature selection, train-test split, model training, hyperparameter optimization, model evaluation, SHAP explainability, and results summary. The design prioritizes accessibility and clinical transparency by excluding ultrasound imaging features and integrating SHAP throughout.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug048

Fig 1 Random Forest Machine Learning Framework with SHAP Explainability

➢ *Dataset*

The dataset was sourced from the Kaggle platform (published by Prasoon Kottarathil) and comprised records from 541 women across 10 hospitals in Kerala, India, with 177 PCOS-positive and 364 non-PCOS cases. The original dataset contained 41 attributes describing physical, clinical, and metabolic parameters.

➢ *Data Preprocessing*

A leakage-free preprocessing pipeline was implemented strictly before data splitting to ensure data integrity. Missing values in continuous variables were imputed using the median, while categorical variables were addressed through mode imputation. Menstrual cycle regularity was recoded into a binary feature, and all categorical string variables (Yes/No) were label-encoded as binary integers (1/0). Duplicate records and irrelevant attributes were removed. To align the model with a non-imaging diagnostic objective, ultrasound-derived features were excluded, reducing the feature space from 41 to 36.

➢ *Feature Selection*

SelectKBest with the ANOVA F-test (f_classif, k=15) was applied to select the most statistically significant features, reducing dimensionality from 36 to 15. This filter-based approach evaluates each feature's individual relationship with the PCOS target variable, reducing noise and mitigating overfitting risk. The selected features are presented in Table 1.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug048

Table 1 Features Selected for Model Training

Variable Category	Features
Demographic	Age (yrs), Marriage Status (yrs)
Anthropometric	Weight (Kg), BMI, Hip (inch), Waist (inch)
Clinical	Weight Gain (Y/N), Hair Growth (Y/N), Skin Darkening (Y/N), Hair Loss (Y/N), Pimples (Y/N), Cycle Length (days), Cycle (R/I)
Biochemical	AMH (ng/mL)
Lifestyle	Fast Food (Y/N)
Target Variable	PCOS Status (PCOS/Non-PCOS)

➢ *Train-Test Split*

The dataset was divided using a stratified 80/20 train-test split, yielding a training set of 432 records and a test set of 109 records. Stratification ensured proportional class representation across both partitions, enabling reliable evaluation of model generalization on unseen data.

➢ *Model Development and Training*

The Random Forest (RF) classifier was selected due to its robustness, resistance to overfitting, and effectiveness with imbalanced clinical data. RF constructs an ensemble of decision trees using bootstrap aggregation (Bagging), with random feature selection at each node to reduce inter-tree correlation [17]. The best split at each node is determined using the Gini Impurity criterion (Eq. 1), and the final classification is obtained by majority voting across N decision trees (Eq. 2).

$$Gini(t) = 1 - Sum(pk^2) \quad \text{(Eq. 1)}$$

Where pk is the probability of a sample belonging to class k at node t, K=2 (PCOS or Non-PCOS).

$$y\text{-}hat = mode\{T1(x), T2(x), ..., TN(x)\} \quad \text{(Eq. 2)}$$

➢ *Hyperparameter Optimization*

The RF model was optimized using RandomizedSearchCV (75 iterations) combined with RepeatedStratifiedKFold cross-validation (5 splits, 2 repeats). This approach provides more robust hyperparameter estimates than standard 5-fold cross-validation, particularly for imbalanced datasets. Class weights were balanced to address class imbalance. A custom probability threshold of 0.46 was adopted to optimize the sensitivity-specificity trade-off, ensuring high detection of true PCOS cases while minimizing false positives.

➢ *SHAP Explainability Module*

To address the black-box limitation of the random forest model, SHAP was applied. SHAP assigns each feature a contribution value (Shapley value) representing its mean marginal contribution to the model prediction across all possible feature subsets [11]. The Shapley value for feature i is computed as (Eq. 3):

$$phi\_i = SUM\ [|S|!(M-|S|-1)!/M!] \times [f(xS+\{i\}) - f(xS)] \quad \text{(Eq.3)}$$

Where S is a subset of features not containing i, M is the total number of features, f(xS) is the model prediction using features in S, and f(xS+{i}) is the prediction with feature i added. SHAP summary plots and mean absolute SHAP value bar charts were generated to visualize global feature importance.

➢ *Model Evaluation Metrics*

Performance was assessed on the hold-out test set using:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad \text{(Eq. 4)}$$

$$Precision = TP / (TP + FP) \quad \text{(Eq. 5)}$$

$$Recall = TP / (TP + FN) \quad \text{(Eq. 6)}$$

$$F1\text{-}Score = 2 \times (Precision \times Recall)/(Precision + Recall) \quad \text{(Eq. 7)}$$

$$Specificity = TN / (TN + FP) \quad \text{(Eq. 8)}$$

The ROC-AUC was computed from the area under the Receiver Operating Characteristic curve, plotting the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) across all thresholds.

IV. RESULTS AND DISCUSSION

➢ *Implementation Environment*

The model was implemented in Python using Google Colab as the development environment. Key libraries included scikit-learn for model training, hyperparameter optimization, and evaluation; pandas and numpy for data manipulation; matplotlib and seaborn for visualization; and the SHAP library for explainability. The dataset was loaded from the Kaggle platform and all preprocessing, feature selection, training, and evaluation steps were executed within this pipeline.

➢ *Model Performance*

The optimized Random Forest model was evaluated on the held-out test set (n=109). The performance metrics are summarized in Table 2 and visualized in Figure 6. The model achieved 92% accuracy, indicating strong overall classification capability using exclusively non-imaging clinical and biochemical data.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug048

Table 2 Performance Metrics of the Random Forest Model

Metric	Value
Accuracy	92.0%
Precision	91.0%
Recall (Sensitivity)	83.0%
F1-Score	87.0%
Specificity	95.8%
ROC-AUC	0.954

The precision of 91% reflects minimal false positive classifications. The 83% recall demonstrates successful identification of the majority of true PCOS cases using only non-imaging features, which is significant for clinical accessibility. The F1-score of 87% confirms an effective balance between precision and recall. The specificity of 95.8% ensures correct classification of non-PCOS cases, limiting unnecessary follow-up. The ROC-AUC of 0.954 indicates excellent discriminatory power across all decision thresholds.

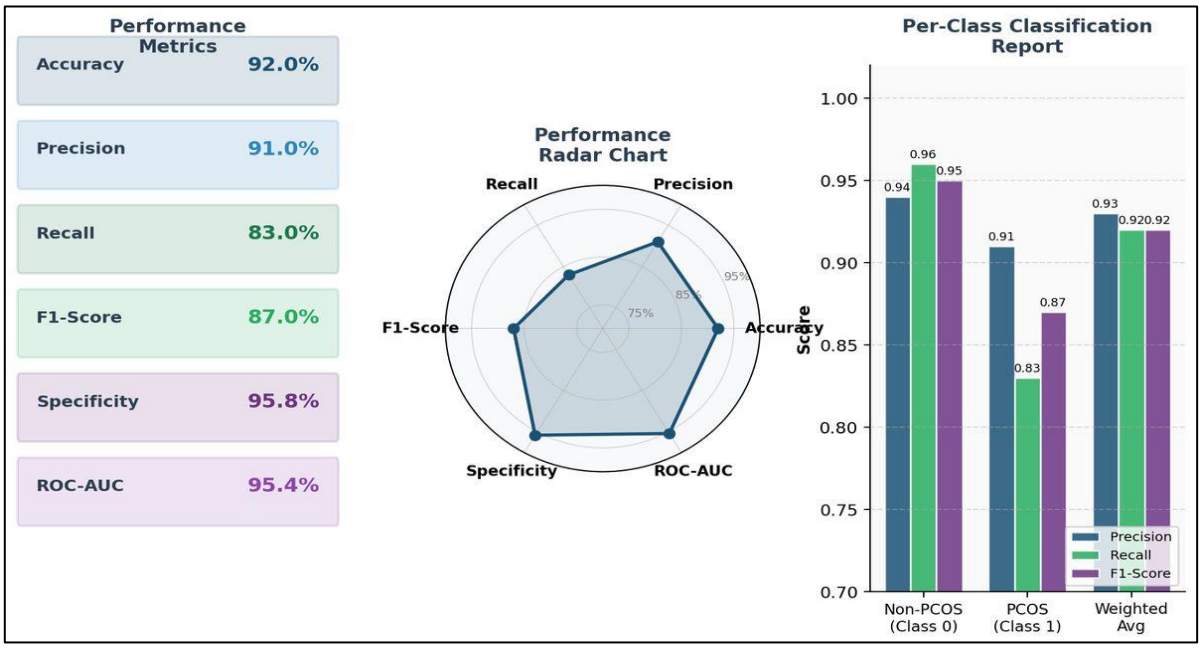

Fig 2 Model Performance Evaluation - Metrics Summary, Radar Chart, and Per-Class Classification Report

➢ *Confusion Matrix Analysis*

The confusion matrix (Figure 2) provides a detailed breakdown of the model's predictions. The RF model correctly classified 70 true negatives (non-PCOS) and 30 true positives (PCOS) on the test set. Only 3 false positives and 6 false negatives were recorded at the applied threshold of 0.46. The low false negative rate is particularly significant clinically, as missed PCOS diagnoses delay treatment and may lead to long-term complications including infertility and metabolic comorbidities.

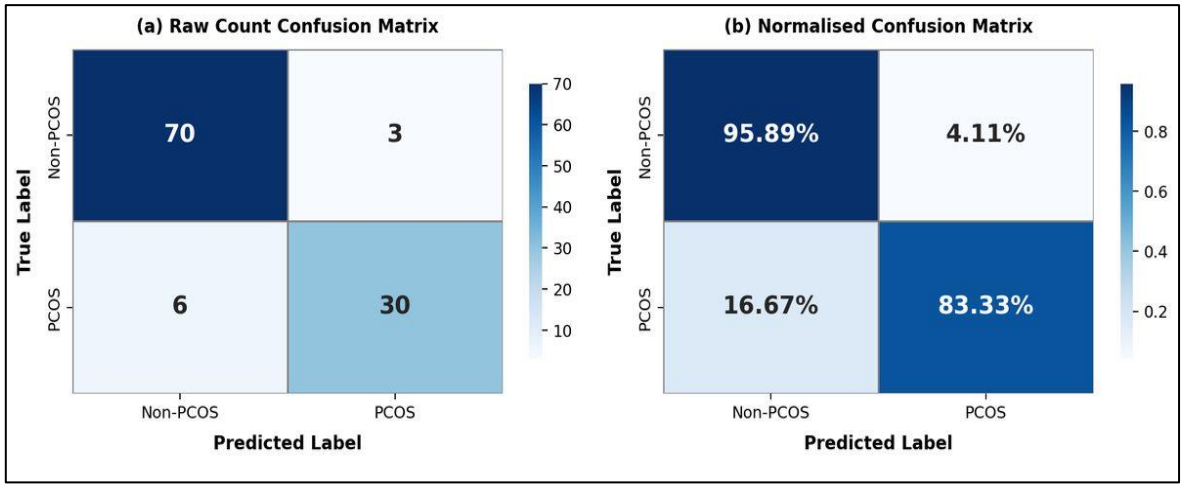

Fig 3 Confusion Matrices for the PCOS Random Forest Model - (a) Raw Count and (b) Normalized

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug048

➢ *ROC Curve Analysis*

The ROC curve (Figure 3) demonstrated excellent discriminatory performance with an AUC of 0.954, confirming that the RF model reliably differentiates PCOS-positive from PCOS-negative cases using only clinical and biochemical features. The operating point at threshold 0.46 achieves 83% sensitivity and 95.8% specificity, as indicated on the ROC curve. This validates the suitability of non-imaging data for building robust PCOS diagnostic models applicable in resource-constrained clinical environments.

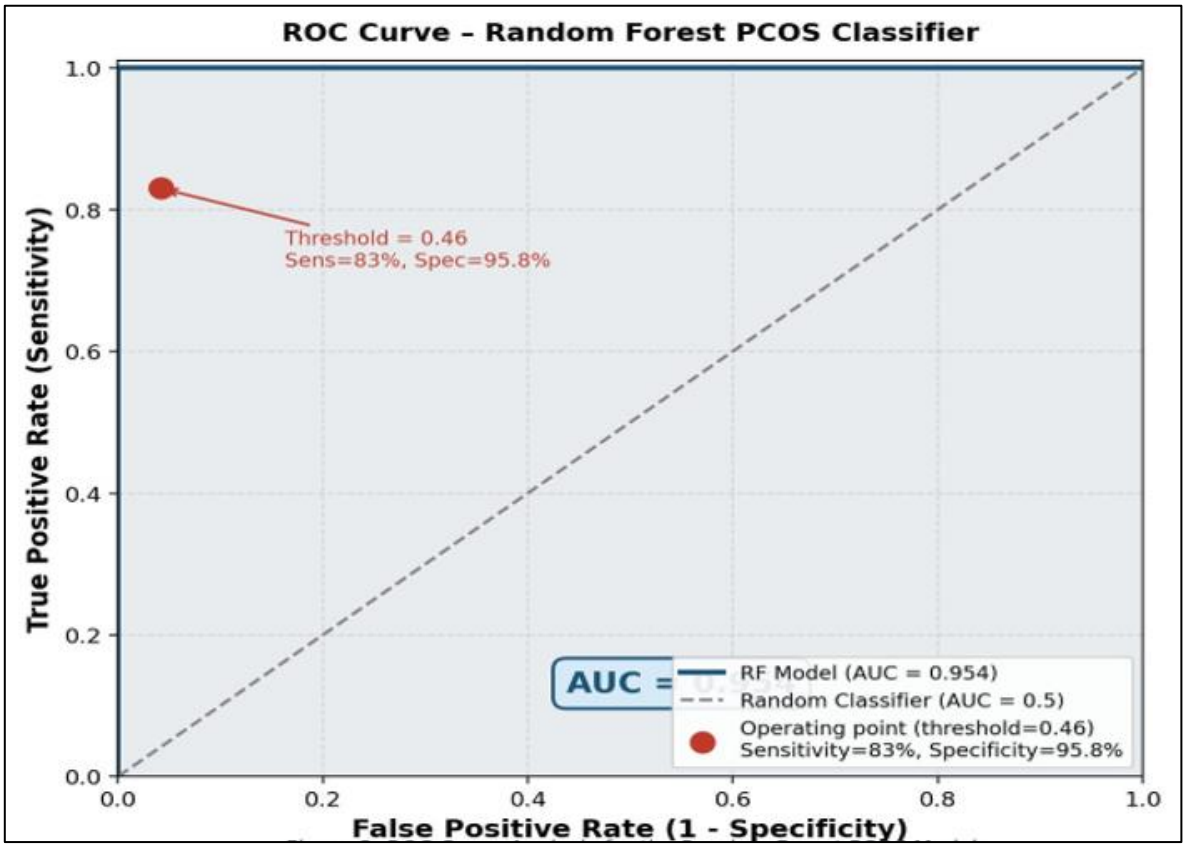

Fig 4 ROC Curve Analysis for the Random Forest PCOS Classifier (AUC = 0.954)

➢ *SHAP Analysis and Model Explainability*

SHAP analysis provided global and local interpretability for the RF model. The SHAP summary beeswarm plot (Figure 4) illustrates the magnitude and direction of each feature's contribution across all test samples. Red data points indicate high feature values and their associated impact on PCOS prediction, while blue points indicate low feature values. The top three most influential features by mean absolute SHAP value (Figure 5) were: (i) weight gain (0.086), (ii) skin darkening (0.078), and (iii) AMH levels (0.067).

These findings are biologically consistent: weight gain promotes hyperinsulinemia and androgen excess; skin darkening (acanthosis nigricans) is a hallmark of insulin resistance; and elevated AMH is a well-validated biomarker of ovarian dysfunction and hyperandrogenism in PCOS [1][18]. These SHAP-identified predictors align strongly with the Rotterdam Criteria and corroborate findings from [3] Cicek and Kucukakcali (2024) and [18], reinforcing the clinical validity of the model's predictive logic.

The SHAP waterfall and feature comparison plots (Figure 7) further illustrate how individual features contribute positively or negatively to specific PCOS predictions, confirming that higher weight gain, skin darkening, AMH, and hair growth values consistently increase PCOS prediction probability across the test cohort.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug048

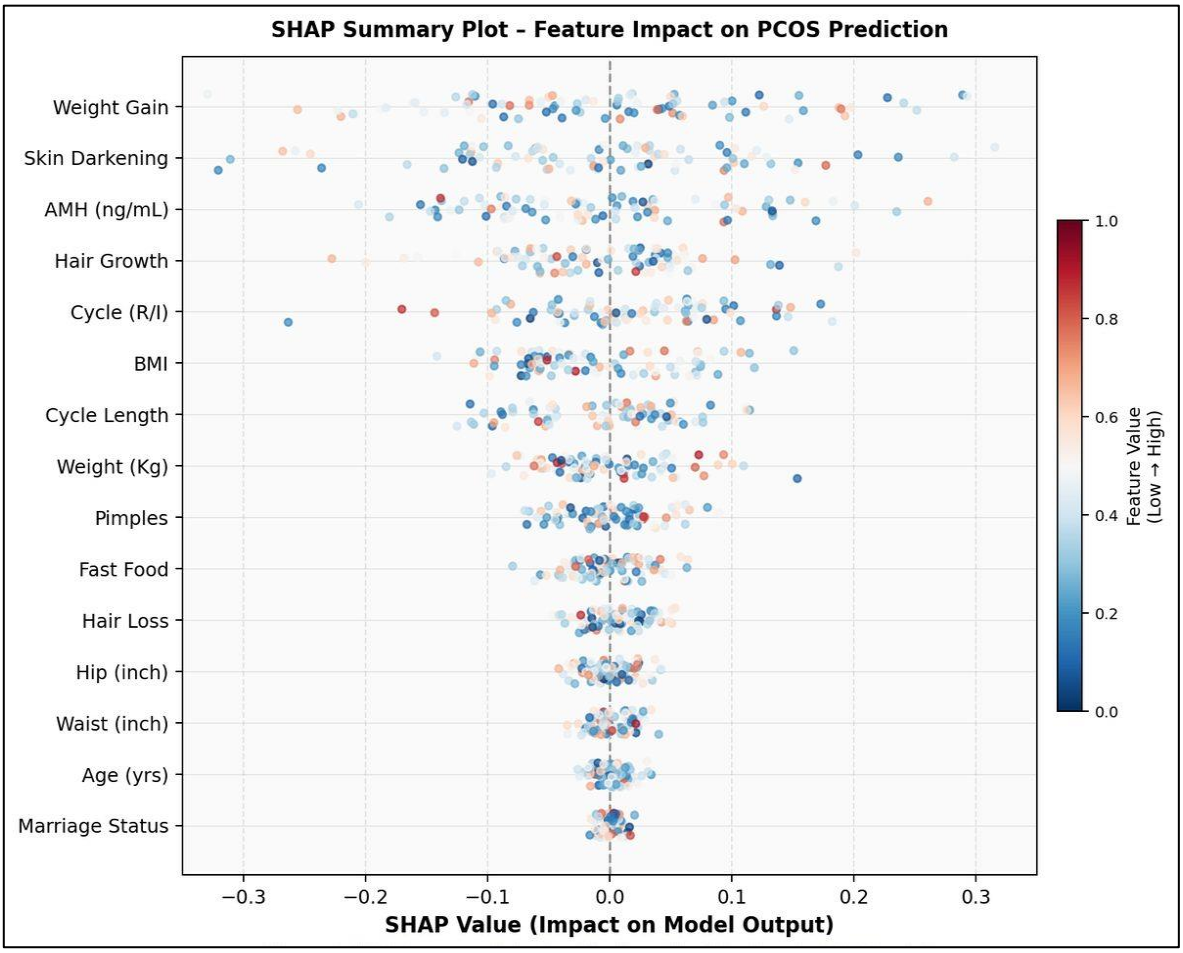

Fig 5 SHAP Summary Beeswarm Plot - Contribution of Each Feature to PCOS Prediction

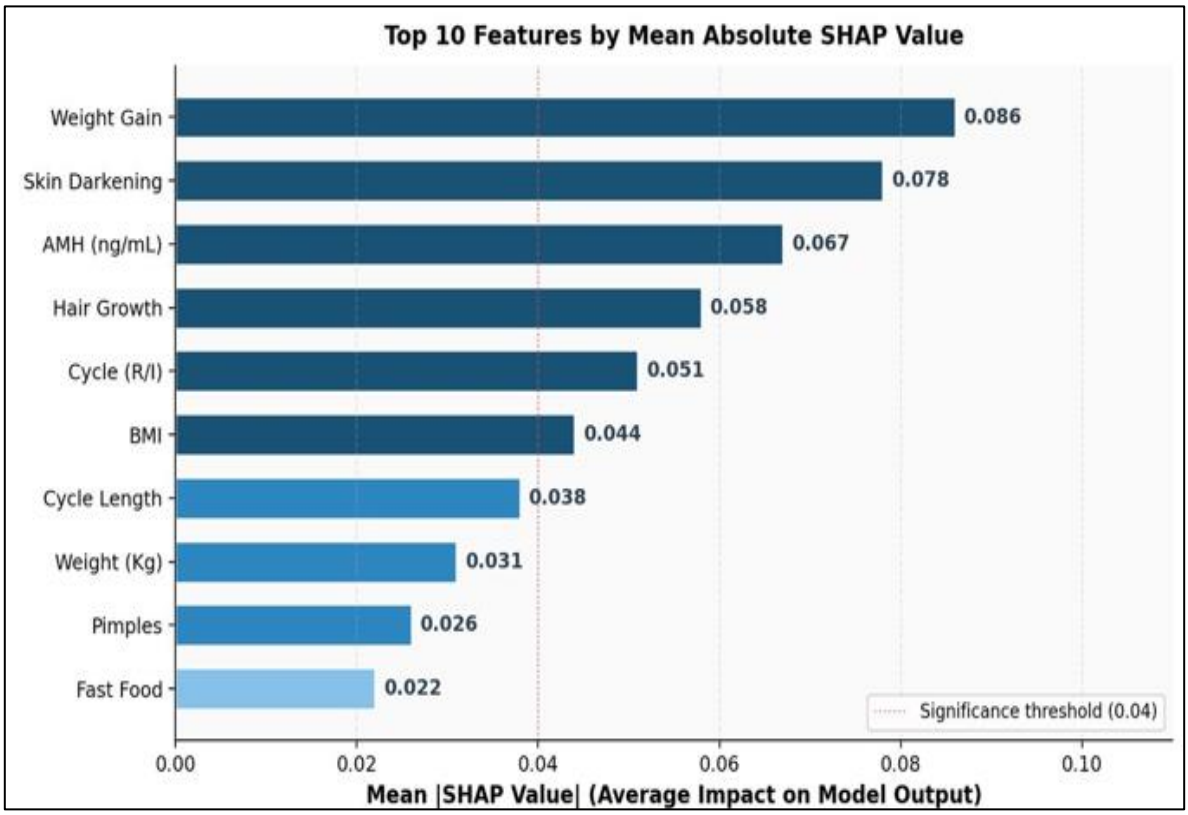

Fig 6 Top 10 Features Ranked by Mean Absolute SHAP Value

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug048

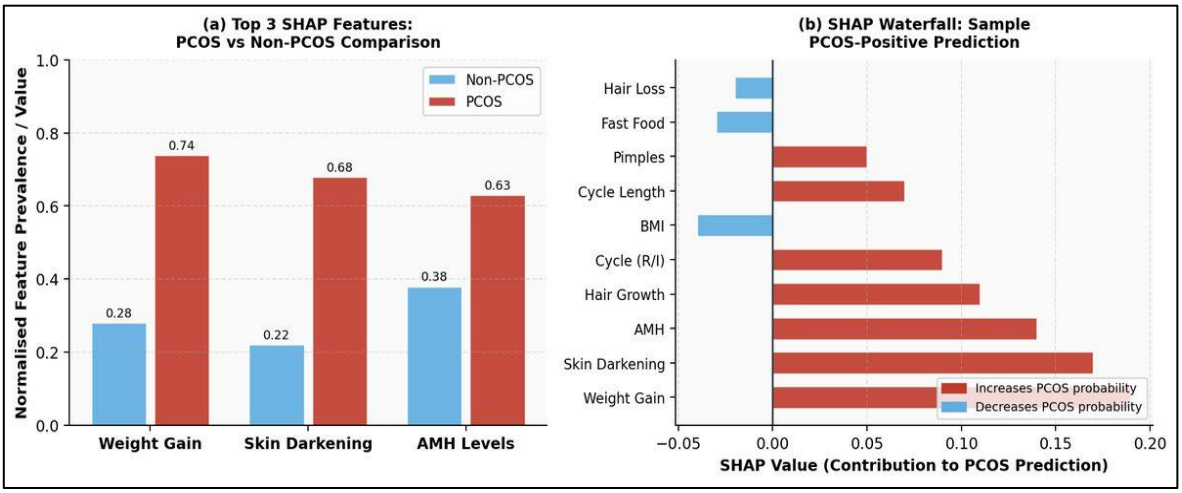

Fig 7 SHAP Feature Analysis - Class Comparison (Left) and Sample PCOS-Positive Waterfall Plot (Right)

➢ *Comparative Analysis*

Table 3 presents a performance comparison with comparable non-imaging studies.

Table 3 Comparative Analysis of Model Performance with Existing Non-Imaging Studies

Study	Method	Accuracy	Precision	Recall	Specificity	F1	AUC
[19]	LSTM & Voting	72.2%	—	—	—	—	0.722
[20]	RF & FFNN	82.0%	85.0%	80.0%	—	—	0.820
[12]	RF + SHAP	86.0%	—	82.0%	91.0%	—	—
This Study	**RF + SHAP**	**92.0%**	**91.0%**	**83.0%**	**95.8%**	**87.0%**	**0.954**

The proposed model outperforms all comparable non-imaging approaches across all reported metrics. The 95.8% specificity and 0.954 AUC substantially exceed those of [12] (91% specificity), while achieving higher accuracy (92% vs. 86%). Unlike [12], which required invasive follicular fluid metabolomics from a limited cohort (n=72), this study employed routinely available clinical and biochemical markers from a larger dataset (n=541), greatly enhancing clinical accessibility and generalizability.

V. CONCLUSION

This study demonstrated that a refined, explainable Random Forest model trained exclusively on non-imaging clinical and biochemical data achieved excellent PCOS diagnostic performance: 92% accuracy, 91% precision, 83% recall, 87% F1-score, 95.8% specificity, and ROC-AUC of 0.954. SHAP explainability confirmed that weight gain, skin darkening, and AMH levels are the most influential predictors, in alignment with established Rotterdam diagnostic criteria. The model's high performance using accessible, non-invasive features makes it particularly suited for screening and early detection in low-resource healthcare settings, bridging the diagnostic gap for women in underserved communities. Future work should focus on external validation using diverse multi-center datasets, prospective clinical evaluation, and development of a mobile or web-based clinical decision support tool. Healthcare policy should consider integrating explainable ML-based screening into routine PCOS assessment workflows to improve early diagnosis, reduce misdiagnosis, and promote transparency in clinical decision-making.

ACKNOWLEDGMENT

The authors acknowledge the contributions of Prasoon Kottarathil for the public availability of the PCOS dataset on the Kaggle platform and Nile University of Nigeria for institutional support.

REFERENCES

[1]. M. Agirsoy, and M.A. Oehlschlaeger, “A machine learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features,” Scientific Reports, 2025, 15(1): 33638.

[2]. World Health Organization, “Polycystic ovary syndrome”. Available at: https://www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome, 2025.

[3]. I. Cicek, and Z. Kucukakcali, “Global interpretation of machine learning models in predicting polycystic ovary syndrome with the explainable artificial intelligence method SHAP,” Medicine Science | International Medical Journal, 2025, 13(3): 568-574.

[4]. S. Mane, R. Aghav, and S. Karad, “PCOS Management: Integrating Contemporary Medicine and Lifestyle Interventions,” Himalayan Journal of Health Sciences, 2025, 10(1): 33-35.

[5]. H.J. Teede, C.T. Tay, J. Laven, A. Dokras, L.J. Moran, T.T. Piltonen, M.F. Costello, J. Boivin, L.M. Redman, J.A. Boyle, R.J. Norman, A. Mousa, and A.E. Joham, “Recommendations from the 2023 International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug048

Syndrome," Fertility and Sterility, 2025, 120(4): 767-793.

[6]. M. Li, Z. He, L. Shi, M. Lin., M. Li, Y. Cheng, H. Liu, L. Xue, Said, K. S., Yusuf, M., Galadanci, H. S. and L. Nie, "Intelligent detection for Polycystic Ovary Syndrome (PCOS): Taxonomy, datasets and detection tools," Computational and Structural Biotechnology Journal, 2025, 27: 1578-1599.

[7]. M. Ahsan, S.A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," Healthcare, 2022, 10(3): 541.

[8]. G.U. Mahesswari, and P.U. Maheswari, "SmartScanPCOS: A feature-driven approach to cutting-edge prediction of Polycystic Ovary Syndrome using Machine Learning and Explainable Artificial Intelligence," Heliyon, 2024, 10(20): e39205.

[9]. B.J. Chelliah, S.K. Gahra, and A. Senthilselvi, "Enhancing PCOS prediction using machine learning and explainable AI," International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT) IEEE. 2024, pp. 1-2.

[10]. V. Sullivan and K. Weger, "Transparency and Explainability in AI-Assisted Decision Making. Proceedings of the Human Factors and Ergonomics Society Annual Meeting," 2025, 69(1): 1827-1833.

[11]. R. Alkhanbouli, H. Matar Abdulla Almadhaani, F. Alhosani and M. Simsekler, "The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions," BMC Medical Informatics and Decision Making, 2025, 25: 110.

[12]. S. Yasar, "Machine Learning-Driven Metabolomic Biomarker Discovery for PCOS: An Interpretable Approach Using Random Forest and SHAP," Medical Records, 2025, 7(3): 763-767.

[13]. N. Faris, and F. Miften, "An intelligence model for detection of PCOS based on k-means coupled with LS-SVM," Concurrency and Computation: Practice and Experience, 2022, 34: e7139.

[14]. Z. Zad, V.S. Jiang, A.T. Wolf, T. Wang, J.J. Cheng, I.C. Paschalidis, and S. Mahalingaiah, "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," Frontiers in Endocrinology, 2024, 15: 1298628.

[15]. F.J. Barrera, E.D. Brown, A. Rojo, J. Obeso, H. Plata, E.P. Lincango, N. Terry, R. Rodriguez-Gutierrez, J.E. Hall, and S. Shekhar, "Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: A systematic review," Frontiers in Endocrinology, 2023, 14: 1106625.

[16]. M. Ghaderzadeh, A. Garavand, and C. Salehnasab, "Artificial intelligence in polycystic ovary syndrome: A systematic review of diagnostic and predictive applications," BMC Medical Informatics and Decision Making, 2025, 25(1): 427.

[17]. H. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," Babylonian Journal of Machine Learning, 2024, pp 69-79.

[18]. B. Panjwani, J. Yadav, V. Mohan, N. Agarwal, and S. Agarwal, "Optimized Machine Learning for the Early Detection of Polycystic Ovary Syndrome in Women," Sensors, 2025, 25(4): 1166.

[19]. J. Lim, J. Li, X. Feng, L. Feng, Y. Xia, X. Xiao, Y. Wang, and Z. Xu, "Machine learning classification of polycystic ovary syndrome based on radial pulse wave analysis. BMC Complementary Medicine and Therapies, 2023, 23(1): 409.

[20]. H. Narni, V. Ananthasetty, and S. Jilani, "Development of a Preliminary Screening Tool for Predicting Polycystic Ovarian Syndrome using Machine Learning and Deep Learning Models with Non-Invasive Qualitative Features: A Case-control Study," Journal of Clinical and Diagnostic Research, 2024, 18(12): KC06-KC10.