PCOS Care: A Machine Learning-Based Web Application for Early Risk Prediction of Polycystic Ovary Syndrome

Rachel Cooray
School of Computer Science and Engineering
University of Westminster
London, UK
rachelcooray@gmail.com
w1956444@my.westminster.ac.uk

Malarvizhi Kaniappan Chinnathai
School of Computer Science and Engineering
University of Westminster
London, UK
M.Kaniappanchinnathai@westminster.ac.uk
https://orcid.org/0000-0001-7044-3120

Salma Chahed
School of Computer Science and Engineering
University of Westminster
London, UK
S.Chahed@westminster.ac.uk
https://orcid.org/0000-0001-7744-5989

Abstract*— Polycystic Ovary Syndrome (PCOS) is a common but underdiagnosed hormonal disorder affecting women of reproductive age, with delayed diagnosis contributing to significant long-term health complications. Although various diagnostic methods exist, there remains a lack of accessible, predictive, and user-centered tools that support its early screening and risk assessment, especially for underserved populations. To address this gap, this paper presents PCOS Care, a Machine Learning (ML) based web application designed to provide early risk prediction of PCOS through a publicly available user-friendly screening. The system utilizes two levels of risk prediction, a simple model based on lifestyle and physical indicators, and an enhanced model incorporating hormonal test data. Models were trained on real-world data, publicly available on Kaggle, using supervised learning classification algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM) and Gradient Boosting with hyperparameter tuning and 10-fold-cross-validation. The enhanced model achieved the best performance with Random Forest, reporting 93.75% recall and 92.66% accuracy. The simple model also demonstrated strong potential, with 83.33% recall and 81.82% accuracy with Logistic Regression. Feature selection was performed using statistical methods such as analysis of variance (ANOVA), chi-square, correlation analysis, and mutual information to select the most relevant predictors. The application integrates a Streamlit-based frontend with a Flask -backend and is deployed on Render using a continuous integration and continuous deployment (CI/CD) pipeline. The usability of the system was tested with 32 potential users, and the results showed high acceptance, ease of navigation and system reliability. Conclusively, PCOS Care illustrates the benefits of ML in healthcare by enabling preliminary risk assessments and empowering users with a digital tool for informed health decisions.

***Keywords*— Polycystic Ovary Syndrome, Machine Learning, Reproductive Health, Risk Prediction, Medical Diagnosis, SMOTE, Streamlit, Flask API, Logistic Regression, Healthcare Technology**

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a widespread hormonal disorder affecting up to 10% of women of reproductive age worldwide [1]. This disorder is associated with reproductive morbidity, increased risk for endometrial cancer, and increased risk of metabolic and cardiovascular diseases [2-3]. Undiagnosed PCOS can also lead to serious long-term health consequences, including infertility [1], obesity, insulin resistance, type 2 diabetes [4] and mental health disorders such as anxiety and depression [5].

Despite its high prevalence and impact on women's health, nearly 70% of PCOS cases remain undiagnosed [5], primarily due to complex diagnostic procedures, limited awareness, and inadequate access to specialized healthcare. Many women with PCOS report being ignored or dismissed by healthcare professionals, often facing years-long delays in diagnosis and having to research their condition independently [6]. Early detection and effective recognition of PCOS is crucial to enhance the individual's quality of life [7].

Traditional diagnosis typically involves a combination of hormonal tests, ultrasound scans, and physical assessments, often requiring multiple clinical visits. This process can be inaccessible, time-consuming, and intimidating, especially for individuals in underserved regions or with cultural stigmas surrounding reproductive health. Moreover, the lack of information and support from healthcare professionals leads to the reliance on online resources and increased stress [3]. Artificial Intelligence (AI) is increasingly used to diagnose complex health issues and AI-based diagnostic tools can support PCOS diagnosis [7]. However, there is a lack of research on the development of AI-based diagnostic tools to support early detection of PCOS.

To address this healthcare gap, this paper presents PCOS Care, a web-based application that uses machine learning (ML) techniques to offer a preliminary risk assessment for PCOS. The system enables users to input clinical and lifestyle information to receive real-time predictions in a user-friendly way and without the need to share or store personal data. To address the challenges associated with limited awareness and to facilitate more effective interventions, PCOS Care also provides reliable health information and visual feedback to raise awareness and encourage early medical consultation.

The main contributions of this research are:

- Statistical feature selection and hyperparameter tuning on real-world dataset using two levels of risk prediction algorithms.
- An interactive web application built using Streamlit and Flask, deployed on Render with continuous integration and continuous deployment (CI/CD).

- A framework to enable preliminary risk assessment and early detection of PCOS using Artificial Intelligence.

This paper also evaluates the broader implications of underdiagnosed PCOS on individual health and healthcare systems. Delayed or missed diagnoses can lead to worsened health outcomes for patients and increased long-term burdens on healthcare infrastructure due to untreated metabolic, reproductive, and other complications.

II. Literature Review

A. *Importance of Early PCOS Detection*

PCOS has emerged as one of the most underdiagnosed hormonal disorders globally. Studies estimate that 5–10% of women of reproductive age are affected, yet approximately 70% remain undiagnosed or misdiagnosed, often due to varied symptom expression and overlapping indicators with other health conditions [5].

Delayed diagnosis can significantly increase the risk of infertility, type 2 diabetes, cardiovascular disease, and mental health disorders, contributing to a substantial burden on healthcare systems as well [8]. This underlines the critical need for accessible, early screening tools that can prompt timely medical intervention.

B. *Existing Applications and Limitations*

Current digital solutions related to PCOS largely focus on symptom tracking or general reproductive health management. Applications such as Flo provide menstrual and symptom tracking features but lack predictive capabilities or medical relevance in terms of risk estimation [9]. The National Health Service (NHS) PCOS information website, while reliable, is purely informational and lacks interactivity, personalized risk analysis, or actionable insights tailored to individual health profiles [10]. There is a clear gap in digital healthcare tools that combine ML with a user-focused interface and responsible data handling to support early PCOS risk identification.

C. *Machine Learning for PCOS Prediction*

Several academic studies have demonstrated the potential of ML in PCOS detection. As shown in [4], a performance comparison of classifiers indicated that Support Vector Machine (SVM) with a linear kernel achieved the best balance of precision (93.67%), accuracy (91.6%), and recall (80.6%), highlighting its suitability for PCOS detection. In another study [11], 18 features were evaluated, primarily based on lifestyle and dietary habits, using multiple classifiers. The Naïve Bayes algorithm achieved the highest accuracy of 97.65% on a dataset of 119 samples via the holdout method, demonstrating the effectiveness of probabilistic approaches. In [12], two-sample t-test to filter nine clinical and metabolic features from a dataset of 200 women. The study compared logistic regression and Bayesian classification, with Bayesian outperforming logistic regression by achieving an accuracy of 93.93% using 3-fold cross-validation. Similarly, a prediction system named i-HOPE used 23 refined features via Principal Component Analysis. Among various models, the Random Forest Classifier yielded the highest accuracy at 89% [13]. In [14], Gradient Boosting, Random Forest, Logistic Regression, and a hybrid Random Forest with Logistic Regression (RFLR) were used to predict PCOS risk. The RFLR method emerged as the most effective, achieving 91.01% accuracy with a recall of 90%. In [5], multiple ML models were trained, and Random Forest Classifier was found to be the most effective, attaining 83.48% accuracy with a reduced feature set of 10, making the model more practical for clinical use [5].

Mutual Information-based feature selection was used in [8] to compare classifiers including Logistic Regression, Decision Tree, AdaBoost, Random Forest, and SVM. Both AdaBoost and Random Forest achieved the highest classification accuracy of 94%, further confirming the reliability of ensemble methods in this domain. A correlation-based feature selection approach was used on a publicly available clinical dataset followed by classification using a wide range of models including XGBoost, CatBoost, and various discriminant analysis techniques. Random Forest outperformed other classifiers with an accuracy of 93.25% [1].

D. *Feature Selection in PCOS Prediction*

Feature selection plays a pivotal role in the development of reliable and interpretable ML models, particularly in the healthcare domain where data dimensionality and clinical relevance are critical. Selecting the correct set of features not only improves model performance but also enhances its generalizability and reduces the risk of overfitting. It is important to note that a lack of proper feature selection could result in models that incorporate noisy, weakly correlated, or redundant variables, subsequently degrading the performance and trustworthiness of ML-based clinical applications.

A univariate filter-based approach was used to rank and select the 10 most informative features from an initial pool of 44, significantly improving classification accuracy while maintaining model simplicity [5]. The work in [4] combined expert input from an obstetrician-gynecologist with the Analysis of Variance (ANOVA) statistical test to identify medically relevant features, ensuring clinical validity alongside statistical strength. Across several studies, statistical methods such as Pearson correlation [1], ANOVA, and chi-square tests[15-16] have been used to determine the most significant predictors of PCOS.

E. *Summary*

The literature review reveals a significant gap which is the lack of deployable, user-validated, and interpretable ML systems for individualized PCOS risk prediction. While existing models show promise, they largely remain in experimental settings, lacking real-world deployment and usability testing, highlighting the critical need for systems developed with diverse data and rigorous validation protocols.

The proposed research builds upon prior work by systematically applying statistical tests to identify the most relevant features. This approach not only enhances model performance and interpretability but also aligns with the clinical goal of creating an explainable and practical decision-support tool for early PCOS risk assessment.

III. System Architecture And Design

The PCOS Care system uses a three-tier architecture, with a frontend, backend Application Programming Interface

(API), and ML layer, designed for scalability, maintainability, and independent development. Figure 1 presents the system architecture design.

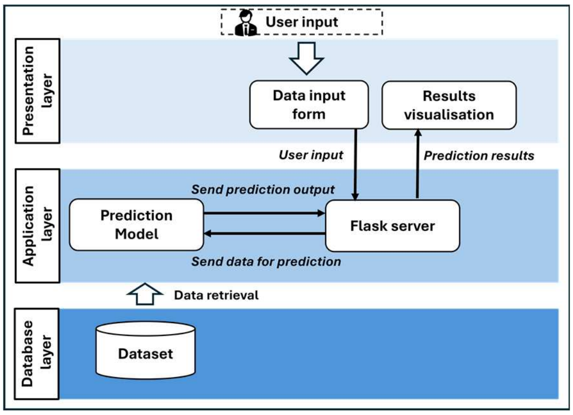

Fig. 1. PCOS Care System Architecture Design

A. System Overview

The system architecture integrates three primary components.

- **Frontend (Presentation Layer):** Built using Streamlit, the web interface allows users to input health-related information and view prediction results. The interface emphasizes simplicity, accessibility, and educational value.
- **Backend (Application Layer):** A Flask-based Representational State Transfer (REST) API processes user input, manages communication between frontend and ML models, and returns prediction outcomes.
- **Model Layer (Database Layer):** Contains serialized ML models trained and stored as Pickle files. These are loaded dynamically during API calls to generate predictions.

B. User Interface (UI) Design

The UI was developed using Streamlit due to its rapid prototyping capabilities and compatibility with Python-based ML workflows. Key features include: (1) form-based input for basic and hormonal health parameters, (2) tooltips and placeholder text for non-technical users, (3) binary prediction output, (4) color-coded output of key health ratios from user data (e.g., green for positive traits, red for negative traits), (5) accessibility features such as appropriate fonts, color contrasting, and descriptive labels to clarify medical terms for non-expert users, and (6) a dashboard with charts giving analysis from the dataset, with insights regarding PCOS. UI wireframes were first prototyped using Figma and then refined based on user feedback and accessibility guidelines.

C. Educational Integration

To address the informational gap around PCOS, the system includes a dedicated educational section. It provides clear, non-technical explanations of symptoms, causes, and lifestyle management strategies. All content was sourced from verified medical references and includes disclaimers that the tool is not a substitute for professional medical diagnosis.

D. Infrastructure and deployment

Deployment was handled using Render.com, which hosts both the frontend (Streamlit) and backend (Flask) as separate services. GitHub was used to implement a CI/CD pipeline, allowing for automatic updates upon repository changes. The ML models were stored on GitHub and loaded by the API at runtime. The following link provides access to the web application:
https://pcos-care-web-app.onrender.com.

IV. Model Development and Training

The core functionality of PCOS Care is driven by supervised ML models trained to predict the likelihood of PCOS based on a patient's clinical and lifestyle data. This section outlines the data preprocessing, feature selection, model training process, and rationale behind the dual-model approach. The complete ML Pipeline is shown in Figure 2.

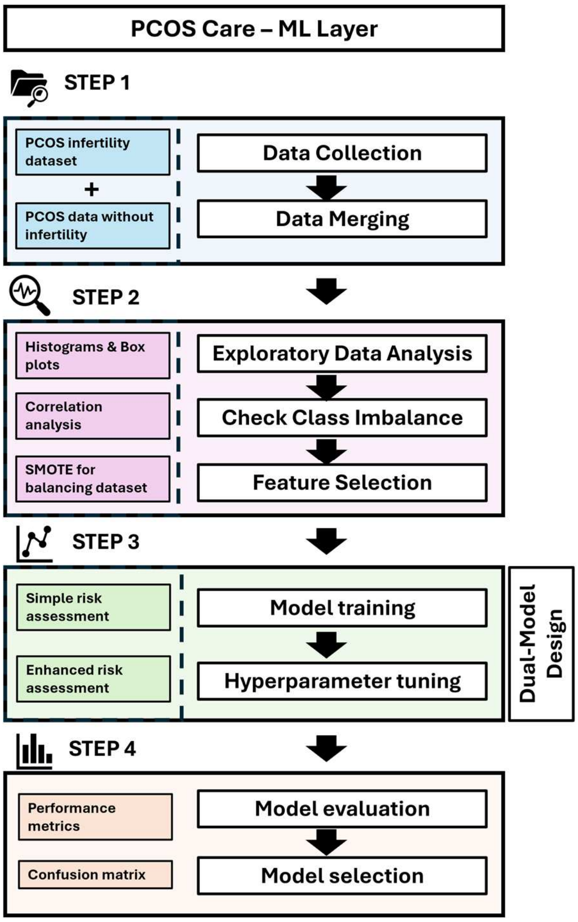

Fig. 2. Steps in the Machine Learning layer of PCOS Care

A. Dataset Description

The dataset used in this study was obtained from Kaggle, titled *"Polycystic Ovary Syndrome (PCOS) Dataset"* [17]. It consists of 541 anonymized patient records collected from 10 hospitals in Kerala, India. Each record includes both numerical (e.g., body mass index (BMI), luteinizing hormone (LH) follicle-stimulating hormone (FSH)) and categorical attributes (e.g., skin darkening, hair growth), along with a binary target indicating PCOS diagnosis. The dataset is imbalanced, with 364 out of 541 records (67.28%) labeled as non-PCOS and only 177 (32.72%) labeled as PCOS.

Therefore, Synthetic Minority Oversampling Technique (SMOTE) is used to balance the dataset. SMOTE is selected because it effectively generates synthetic minority samples through interpolation, improving classifier performance on imbalanced healthcare datasets while minimizing overfitting associated with random oversampling [20].

B. Data Preprocessing

The following steps were implemented as part of data preprocessing: (1) missing value imputation for numerical features (using median values) and categorical features (using mode), (2) Z-score normalization applied to all numerical features to standardize scale and improve model convergence, and (3) binary encoding for categorical variables.

C. Dual-model Design

The proposed methodology enables dual-model prediction. The simple model uses basic lifestyle and physical indicators, including inputs like exercise habits, fast food intake, acne, and skin darkening for predicting PCOS risk. In this case, the users do not need clinical test or scan results. On the other hand, the enhanced model incorporates features from the simple model along with clinical data such as hormone levels (e.g., LH, FSH), which are typically obtained through diagnostic tests. This ensures that the system remains accessible to the general public while offering higher prediction accuracy for users with clinical results.

D. Feature Selection

The dataset originally contained 42 features, encompassing lifestyle, clinical, and hormonal variables. To improve model performance and interpretability, a multi-stage feature selection process was conducted that included manual filtering, domain-driven selection, and statistical evaluation as shown in Figure 3.

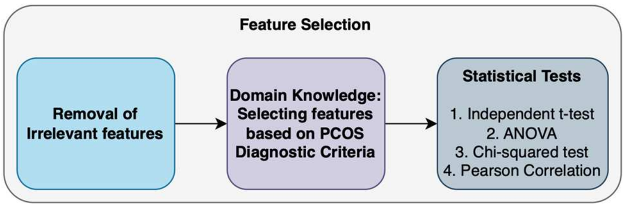

Fig. 3. Feature selection process

1) Removal of Irrelevant or Redundant Features

Initial cleaning involved excluding variables that offered little to no predictive value for PCOS. For example, demographic variable "Marriage Status (Yrs)" was removed as it did not directly contribute to the clinical diagnosis of PCOS and could introduce noise into the model.

2) Domain Knowledge: PCOS Diagnostic Criteria

To ensure clinical relevance, features were aligned with established diagnostic guidelines. According to the Rotterdam Criteria (2003), PCOS diagnosis requires at least two out of the following three: (1) Ovulatory dysfunction (e.g., irregular cycle), (2) Hyperandrogenism (e.g., acne, hirsutism, hair loss), and (3) Polycystic ovaries.

Complementary criteria from the National Institutes of Health (NIH) (1990) and Androgen Excess and PCOS (AE-PCOS) Society (2006) were also reviewed [18]. Based on this domain knowledge, features directly reflecting these aspects, such as FSH, LH, Anti-Müllerian Hormone (AMH), follicle count (both ovaries), and androgen-related symptoms were retained for further analysis.

3) Statistical Tests

Statistical feature selection methods were used to identify significant predictors of PCOS. Features were first screened using independent t-tests, ANOVA, and Chi-squared tests, with a significance level of $\alpha = 0.05$. For numerical features, Pearson correlation was used to assess linear relationships with the binary target variable, and features with higher correlation coefficients were retained to ensure a meaningful association. Mutual information was also calculated to capture potential non-linear dependencies and support the selection process.

4) Model Training

Several classification algorithms were evaluated during model development, including Logistic Regression, SVM, Random Forest, Gradient Boosting, Decision Trees, K-Nearest Neighbors (KNN), and Multi-layer Perceptron (MLP). Each model was trained using a 10-fold cross-validation approach to ensure balanced representation of both classes across all folds. GridSearchCV was used to perform hyperparameter tuning for each algorithm. Sensitivity analysis was conducted using varying train-test splits of 80:20, 85:15, and 90:10 to evaluate model stability across different sample sizes.

V. TESTING

To ensure reliability and usability, the proposed system was subject to comprehensive testing at both technical and user levels. Unit tests using Python's unittest and pytest validated input handling, API routing, and prediction formatting. Postman was used to test JavaScript Object Notation (JSON) payloads, responses, and error handling. Model integration tests confirmed Pickle files loaded and functioned correctly. System testing evaluated performance and compatibility through end-to-end flows and mock sessions. Though formal stress testing was not conducted, user feedback confirmed fast performance. The application was also tested across browsers and devices, meeting Web Content Accessibility Guidelines. Usability testing with 32 participants showed high satisfaction, with positive feedback on navigation, clarity, usefulness and trust.

VI. RESULTS AND EVALUATION

The PCOS Care system was evaluated through model performance metrics, user testing feedback, and qualitative analysis of its utility as a digital health tool. Emphasis was placed on minimizing false negatives (FN), improving usability, and supporting informed health decisions.

A. Model Performance

Multiple ML models were trained and evaluated; the results of the simple prediction model are presented in Table 1. The results of the enhanced prediction model are presented in Table 2.

TABLE I. PERFORMANCE EVALUATION OF THE SIMPLE MODEL

Split	Algorithm	Accuracy	Recall	Precision	F1-Score	FN
80/20	Logistic Regression	0.8073	0.7500	0.7500	0.7200	9
	MLP	0.8257	0.7222	0.7429	0.7324	10
	Gradient Boosting	0.8165	0.6667	0.7500	0.7164	12
	KNN	0.8165	0.6667	0.7500	0.7059	12
	Decision Trees	0.7431	0.6667	0.6000	0.6316	12
	SVM	0.7890	0.6389	0.6970	0.6667	13
	Random Forest	0.8349	0.6111	0.8462	0.7097	14
85/15	Logistic Regression	0.8415	0.8148	0.7333	0.7719	5
	MLP	0.8293	0.7407	0.7407	0.7407	7
	Gradient Boosting	0.8171	0.6667	0.7500	0.7059	9
	Random Forest	0.8293	0.6296	0.8095	0.7083	10
	SVM	0.7439	0.6296	0.6071	0.6182	10
	Decision Trees	0.7561	0.5556	0.6522	0.6000	12
	KNN	0.6951	0.5556	0.5357	0.5455	12
90/10	**Logistic Regression**	**0.8182**	**0.8333**	**0.6818**	**0.7500**	**3**
	Decision Trees	0.8000	0.8333	0.6522	0.7317	3
	MLP	0.7636	0.6111	0.6471	0.6286	7
	Gradient Boosting	0.8364	0.7778	0.7368	0.7568	4
	Random Forest	0.7818	0.6667	0.6667	0.6667	6
	KNN	0.7455	0.6667	0.6000	0.6316	6
	SVM	0.6909	0.5000	0.5294	0.5143	9

The metrics used for evaluation were as follows, in the order of the priority given to them.

1. High value of recall was prioritized as false negatives could lead to users overlooking serious health risks.
2. High value of F1-Score ensures a good balance between precision and recall, especially important in imbalanced datasets.
3. High value of precision reduces false positives (FP), preventing unnecessary stress for users who are not at risk.
4. Low number of False negatives (FN) is crucial to avoid missing actual PCOS cases.
5. High value of Accuracy reflects overall correctness of predictions but is considered secondary to recall in this context.

From Table 1, Logistic Regression model with 90/10 train/test split achieved the best performance with a recall of 0.833. From Table 2, it can be seen that Random Forest with 80/20 split performed best with a recall of 0.9375. It is also interesting to note that the enhanced model outperformed the simple model due to the availability of hormonal features.

TABLE II. PERFORMANCE EVALUATION OF THE ENHANCED MODEL

Split	Algorithm	Accuracy	Recall	Precision	F1-Score	FN
80/20	Logistic Regression	0.9083	0.8421	0.8889	0.8649	4
	Random Forest	**0.9266**	**0.9375**	**0.8333**	**0.8824**	**6**
	Decision Trees	0.8807	0.8108	0.8333	0.8219	6
	Gradient Boosting	0.8991	0.8788	0.8056	0.8406	7
	MLP	0.8716	0.8056	0.8056	0.8056	7
	KNN	0.8349	0.7250	0.8056	0.7632	7
	SVM	0.8532	0.7778	0.7778	0.7778	8
85/15	MLP	0.8780	0.7931	0.8519	0.8214	4
	Random Forest	0.8902	0.8462	0.8148	0.8302	5
	Logistic Regression	0.8415	0.7333	0.8148	0.7719	5
	Gradient Boosting	0.8780	0.8400	0.7778	0.8077	6
	KNN	0.8049	0.6774	0.7778	0.7241	6
	Decision Trees	0.8293	0.7407	0.7407	0.7407	7
	SVM	0.8171	0.7308	0.7037	0.7170	8
90/10	MLP	0.8545	0.7273	0.8889	0.8000	4
	Gradient Boosting	0.9091	0.9333	0.7778	0.8485	4
	KNN	0.8182	0.7000	0.7778	0.7368	4
	Random Forest	0.8727	0.8667	0.7222	0.7879	5
	Logistic Regression	0.7818	0.6500	0.7222	0.6842	5
	Decision Trees	0.8182	0.7857	0.6111	0.6875	7
	SVM	0.7636	0.6667	0.5556	0.6061	8

B. *Usability Evaluation*

A survey (32 participants) was conducted to assess the usability of the proposed system. Accordingly, users found the tool helpful and 93.75% reported that it was intuitive and easy to navigate. It was also pointed out that the application was quick to load. Although 31% experienced form submission errors, all of them agreed that the error handling was effective. Feedback also emphasized the value of the dual-model risk prediction approach for broader accessibility.

C. *Educational and Healthcare Impact*

PCOS Care acts as a preliminary screening tool that raises awareness and encourages early consultation, helping reduce complications like infertility and diabetes. Delayed PCOS diagnosis increases chronic disease burden and strains healthcare systems. By promoting early care-seeking, especially in underserved populations, digital tools like this can help patients as well as ease pressure on healthcare infrastructure and reduce long-term costs associated with undiagnosed cases [19].

VII. Conclusion and Future Work

This study presents PCOS Care, an ML-based clinical decision support tool for early risk prediction of Polycystic Ovary Syndrome. By combining statistical feature selection, explainable models, and a publicly available web interface, the system addresses the gap in personalized, AI-driven screening tools for women's health, particularly for PCOS. The two-level prediction design supports both lifestyle-based and clinically relevant inputs, aligning with principles of patient-centered care. Usability testing confirmed the tool's effectiveness and acceptability. While not a substitute for clinical diagnosis, PCOS Care aids medical decision-making by empowering users to perform preliminary risk assessments and seek timely consultation contributing to early intervention, clinical impact, and reduced healthcare burden. The dataset's limitations highlight the need for training on larger, diverse datasets, and real-world validation with healthcare partners is the next important step.

While PCOS Care successfully demonstrates the feasibility of integrating ML with accessible digital health tools, several areas remain open for future enhancement and validation. Clinical validation through partnerships with healthcare institutions, pilot testing in real-world settings, and enhancing model personalization using explainable AI can be implemented. Developing a mobile version with multilingual functionality for wider accessibility and incorporation of long-term monitoring features to support ongoing health tracking and personalized PCOS risk updates could improve the application.

Acknowledgment

The authors would like to thank the open-source community and Kaggle contributors for the PCOS dataset, the survey participants for their valuable input, and volunteer testers whose feedback helped refine the system.

References

[1] S. Tiwari, L. Kane, D. Koundal, A. Jain, A. Alhudhaif, K. Polat, et al., "SPOSDS: A smart polycystic ovary syndrome diagnostic system using machine learning," Expert Systems with Applications, vol. 203, May 2022, Art. no. 117592, doi: 10.1016/j.eswa.2022.117592.

[2] R. A. Lobo and E. Carmina, "The importance of diagnosing the polycystic ovary syndrome," Ann. Intern. Med., vol. 132, no. 12, pp. 989–993, 2000, doi: 10.7326/0003-4819-132-12-200006200-00010.

[3] J. Tomlinson, J. Pinkney, L. Adams, E. Stenhouse, A. Bendall, O. Corrigan, and G. Letherby, "The diagnosis and lived experience of polycystic ovary syndrome: A qualitative study," *J. Adv. Nurs.*, vol. 73, no. 10, pp. 2318–2326, 2017, doi: 10.1111/jan.13300.

[4] Y. A. Adla, H. S. Bitar, M. S. Assaf, R. A. Saad, D. G. Raydan, J. Nasreddine, et al., "Automated detection of polycystic ovary syndrome using machine learning techniques," in Proceedings of the 2021 Sixth International Conference on Advances in Biomedical Engineering, pp. 1–4, 2021, doi: 10.1109/ICABME53305.2021.9604905.

[5] D. Gandhi, B. Patel, and N. Dave, "PCOS detection using machine learning algorithms," Indian Journal of Natural Science, vol. 14, no. 81, 2024.

[6] "Women with painful condition 'ignored' by doctors," *BBC News*, Jun. 3, 2024. Accessed: Dec. 16, 2025. [Online]. Available: https://www.bbc.co.uk/news/articles/clwwgl98dg0o

[7] P. Verma, P. Maan, R. Gautam, and T. Arora, "Unveiling the role of artificial intelligence (AI) in polycystic ovary syndrome (PCOS) diagnosis: A comprehensive review," *Reproductive Sciences*, vol. 31, no. 10, pp. 2901–2915, 2024, doi: 10.1007/s43032-024-01615-7.

[8] M. Rahman, T. Tuba, M. A. H. Khan, and M. S. Rahman, "Empowering early detection: A web-based machine learning approach for PCOS prediction," *Informatics in Medicine Unlocked*, vol. 47, 2024, doi: 10.1016/j.imu.2024.101500.

[9] "Flo - ovulation calendar, period tracker, and pregnancy app," 2023. Accessed: Oct. 1, 2024. [Online]. Available: https://flo.health/

[10] NHS, "Polycystic ovary syndrome," *NHS*, 2022. Accessed: Oct. 1, 2024. [Online]. Available: https://www.nhs.uk/conditions/polycystic-ovary-syndrome-pcos/ .

[11] V. B. Vikas, B. S. Anuhya, M. Chilla, and S. Sarangi, "A critical study of polycystic ovarian syndrome (PCOS) classification techniques," *International Journal of Computational Engineering and Management*, vol. 21, no. 4, pp. 1–6, Jul. 2018, [Online]. Available: http://www.ijcem.org.

[12] P. Mehrotra, J. Chatterjee, C. Chakraborty, B. Ghosh Dastidar, and S. Ghoshdastidar, "Automated screening of polycystic ovary syndrome using machine learning techniques," *India Conference (INDICON), 2011 Annual IEEE)*, Dec. 2011, doi: 10.1109/INDCON.2011.6139331.

[13] A. Denny, A. Raj, A. Ashok, C. M. Ram and R. George, "i-HOPE: detection and prediction system for Polycystic Ovary Syndrome (PCOS) using machine learning techniques," *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Kochi, India, 2019, pp. 673-678, doi: 10.1109/TENCON.2019.8929674.

[14] S. Bharati, P. Podder, and M. R. Mondal, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms", *2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh*, doi: 10.1109/TENSYMP50017.2020.9230932.

[15] M. S. Khan Inan, R. E. Ulfath, F. I. Alam, F. K. Bappee, and R. Hasan, "Improved sampling and feature selection to support extreme gradient boosting for PCOS diagnosis," *Proceedings of the 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC),* January 2021, pp. 1046–1050, doi: 10.1109/CCWC51732.2021.9375994.

[16] S. Vedpathak and V. Thakre, "PCOcare: PCOS detection and prediction using machine learning algorithms," *Bioscience Biotechnology Research Communications,* vol. 13, pp. 240–244, Dec. 2020, doi: 10.21786/bbrc/13.14/56.

[17] P. Kottarathil, 2021, "Polycystic Ovary Syndrome (PCOS)", Kaggle. [Online]. Available: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos/data

[18] L. Casadei, F. Fanisio, R. P. Sorge, M. Collamarini, E. Piccolo, and E. Piccione, "The diagnosis of PCOS in young infertile women according to different diagnostic criteria: The role of serum anti-Müllerian hormone," *Arch. Gynecol. Obstet.*, vol. 298, no. 1, pp. 207–215, 2018, doi: 10.1007/s00404-018-4803-8.

[19] H. Elmannai, N. El-Rashidy, I. Mashal, M. A. Alohali, S. Farag, S. El-Sappagh, and H. Saleh, "Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence," *Diagnostics (Basel)*, vol. 13, no. 8, p. 1506, 2023, doi: 10.3390/diagnostics13081506.

[20] S. Yadav, "SMOTE in Predictive Modeling: A Comprehensive Evaluation of Synthetic Oversampling for Class Imbalance," *Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci.*, vol. 8, no. 4, pp. 1–9, August. 2020, doi: 10.5281/zenodo.14259555.