Vol. 8, No. 3, June 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

PCOS Classification Using Random Forest, Recursive Feature Elimination, and Explainable AI

Syifa Ayu Salsabila Putri[1], Rona Nisa Sofia Amriza[2]

[1,2]Information System Study Program, Telkom University, Purwokerto Campus, Indonesia

Received:
October 12, 2025
Revised:
April 26, 2026
Accepted:
May 30, 2026
Published:
June 22, 2026

Corresponding Author:
Author Name*:
Rona Nisa Sofia Amriza
Email*:
ronanisa@telkomuniversity.ac.id

DOI:
10.63158/journalisi.v8i3.1603

Abstract. Polycystic Ovary Syndrome (PCOS) is an endocrine-related condition predominantly affecting women during their childbearing years who experience delayed diagnosis due to the limitations of conventional methods that require laboratory tests and imaging procedures that are relatively costly and time-consuming. This study develops a PCOS classification model based on a clinical dataset of 541 patients with 42 clinical attributes using the random forest algorithm with Recursive Feature Elimination (RFE) feature selection and an Explainable AI (XAI) approach. The research pipeline comprised several sequential stages: problem identification, data collection, preprocessing, data splitting, feature selection, model training and testing, evaluation, and SHAP-based explainability analysis. Performance was evaluated using Accuracy, Precision, Recall, and F1-score, and compared between two models, namely RF+CF and RF+RFE, where RF+RFE was identified as the best-performing model. The XAI approach using SHAP (SHapley Additive exPlanations) was applied to identify and explain the contribution of clinical variables to the classification results. The best model, RF+RFE, achieved an accuracy of 92.66%, precision of 93.75%, recall of 83.33%, and F1-score of 88.24%, demonstrating superior performance compared to RF+CF. As this study relies on a single dataset, broader validation across multiple centers is recommended before clinical deployment. This model is intended as a screening-support approach and has not been validated as a clinical diagnostic tool. The findings are anticipated to serve as a foundation for building data-driven early screening tools and clinical decision-making support systems.

Keywords: Clinical Classification, Feature Selection, PCOS Classification, Recursive Feature Elimination, Explainable AI.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

PCOS ranks among the most prevalent hormonal disorders encountered in women within their reproductive years [1],[2][3]. This disorder was first described by Leventhal and Stein in 1935 [2],[4]. Women with PCOS experience hormonal imbalances that cause health problems such as irregular menstrual cycles and difficulty conceiving [5]. PCOS is not only a reproductive disorder, but is also associated with long-term metabolic complications such as type 2 diabetes, hypertension, obesity, and an increased risk of endometrial cancer [1],[2],[4],[6]. Women diagnosed with PCOS frequently present with a constellation of clinical manifestations, including menstrual irregularities, pronounced acne, abnormal hair growth (hirsutism), impaired glucose metabolism, and hyperpigmentation of the skin such as acanthosis nigricans [2],[4],[5],[7]. According to the World Health Organization (WHO, 2026), PCOS is estimated to affect 10–13% of women of reproductive age worldwide, and 70% of women affected are undiagnosed. Global studies report that the proportion of reproductive-age women affected by PCOS ranges from 6% to 21%, with variation attributable to differences in the diagnostic standards applied6. In Indonesia, there are approximately 4–6% cases of PCOS in women of reproductive age and 75% in women with infertility due to anovulation [2].

The diagnosis of PCOS can be quite complicated. Currently, the diagnostic process still relies on the 2003 Rotterdam criteria, which involves complex laboratory hormone tests and ovarian ultrasounds, and requires additional costs [2],[3],[4],[8]. Delayed diagnosis increases the risk of certain complications, while early detection has been shown to significantly detect early long-term incidents [5],[6]. Machine learning (ML) has emerged as an approach that offers great opportunities by improving the effectiveness of detecting key risk factors for various complex diseases through more objective, accurate, and automated analysis of clinical data and images [5],[9]. Within this study, machine learning is primarily employed to stratify disease risk from structured clinical data, offering a more rapid and consistent diagnostic pathway with considerable promise as a non-invasive PCOS screening approach [6],[9],[10]. A growing body of research has demonstrated the considerable capability of machine learning techniques in both classifying and predicting PCOS outcomes, with algorithms including RF, SVM, DT, LR, KNN, and XGBoost having been evaluated across different study settings with differing degrees of success [1],[3],[4]. RF can improve accuracy by combining multiple Decision

Tree, as well as effectively identifying significant factors in the diagnosis of chronic diseases, including PCOS [10],[11]. A comparison of several models conducted in previous studies shows that PCOS classification using RF achieved the highest results with an accuracy of 92.4% at a correlation threshold of 0.8 [5].

The use of RF models on high-dimensional data has the potential to reduce model performance because the large number of features often contains irrelevant and redundant attributes, which can reduce accuracy [12], [13]. While prior studies have applied Random Forest to PCOS classification, most relied on correlation-based or manual feature selection without systematic algorithmic evaluation of feature contribution, and none have explicitly combined RFE-based feature selection with SHAP-based explainability within a unified pipeline [1],[14],[15]. Even though RF-only approaches could gain reasonable accuracy, they are not aimed to eliminate irrelevant features or provide transparent elaborations on the way every clinical variable gives a contribution to the outcome of classification. Therefore, their clinical utility stays limited, particularly related to both of their feature efficiency and interpretability. On the other hand, the proposed RF+RFE+SHAP pipeline deals with these limitations by collaborating the selection of systematic features with reasonable model outputs in a single unified framework. The main obstacles dealed in this study is connected to the selection of the most relevant clinical features through a systemic algorithmic approach, managing class imbalance within the dataset and increasing the sensitivity of the model to identify positive PCOS issue accurately. This study offers the utilization of the Random Forest ML algorithm with Recursive Feature Elimination (RFE) feature selection along with the utilization of Explainable Artificial Intelligence (XAI). Feature selection is demonstrated to measure the inter-correlation among features. The goal is to avoid redundancy and increase the accuracy of model by paying more attention to the learning process on the features that are most relevant to the class label [12],[13],[16],[17]. In this case, XAI is implemented to elaborate the results into more understandable interpretations [18],[19],[20],[21]. Meanwhile, Accuracy, Precision, Recall, and F1-score aimed to facilitate a comprehensive illustration of the classifier's capability to predict [1],[5],[22].

This study intends to fulfill the gap among prior studies by integrating RFE-Based feature selection along with SHAP-Based reasoning into a Random Forest categorization pipeline for PCOS identification by implementing systematic data clinic. The main goal is to

develop a clinically grounded PCOS categorization model which encourages medical workers to mention precise, efficient, and direct interpretation of diagnoses. The main contribution of this study is divided into three aspects, including (1) the improvement of a Random Forest classifier with RFE-based feature selection, (2) a comparative review against Correlation Filter (CF) as a substitution of feature selection approach, and (3) the utilization of SHAP-based XAI to increase the clinical interpretation of model outputs. This improvement is intended to mediate the gap between the requirement for effective, non-invasive, and affordable PCOS diagnosis and traditional approaches which tend to be more pricey and complicated, and to contribute as a fundamental to develop evidence-based systems of clinical decision-making in clinics or hospitals [10],[19], [23].

2. METHODS

This study implemented a quantitative data mining framework to reveal the main variables' contribution to the onset of PCOS systematically. This is not the same with the conservative statistical techniques. Data mining is able to process substantial volumes of data to uncover latent patterns and complicated relationships among inter-variables which would not be identified [9],[12]. This study suggests the use of a structured PCOS categorization framework which include the application of Recursive Feature Elimination (RFE) and SHAP-based Explainable AI (XAI) to create clinical interpretation in a unified Random Forest pipeline.

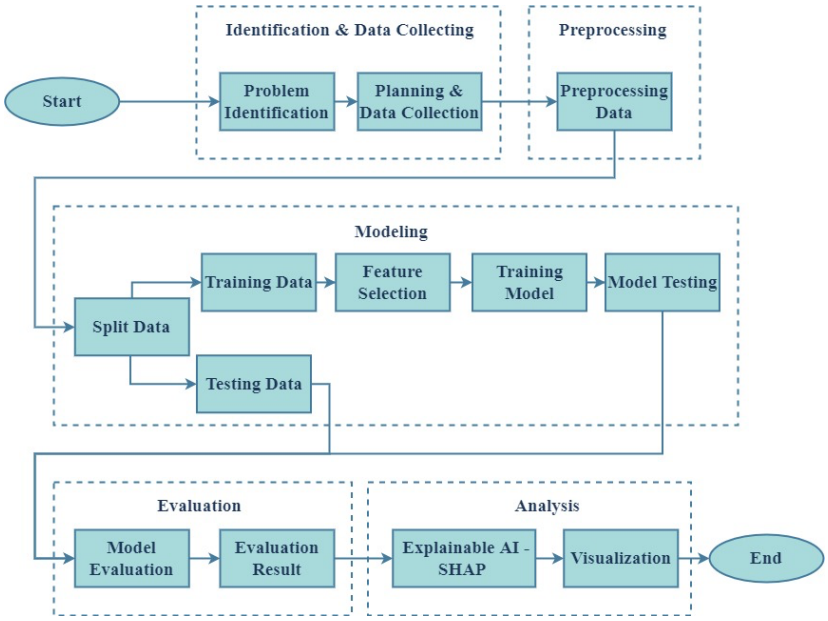

Figure 1. Methodology for classification of PCOS

This study was developed by having a structured research pipeline including some ordered stages, as depicted through Figure 1. These sequences start from Identification of Problem; Planning and Collection of Data; Preprocessing of Data; Data Splitting; Feature Selection; Model Training, Testing, and Evaluation; Explainable AI implementation; and Results Visualization.

2.1. Problem Identification and Data Collection

The problem identification in this study is aimed to collect the primary cases to be assessed. The focus is targeted to recognizing PCOS as an endocrine disease which occurs in women during their reproductive ages with various charateristics of clinic. The identified cases are related to the complexity of deciding the dominating risk factors due to the high number of connected variables. This identification was brought up through the analysis of this study needs, prior data evaluation, and relatable scientific references. These obtain a clear question as the foundation to decide the methods, data collection, and process of the study. As seen from the problem identification, it was revealed that varieties in diagnostic characterization referring to the Rotterdam also contribute to phenotypic heterogeneity and distinctions in characteristics among the populations. This holds the ability to impact the stability along with predictive models' generalization.

The implementation of Random Forest as a classification method has definitely proven well presentation in various PCOS studies [18]. However, this model is still limited in interpretability because of its complicated ensemble model of learning. These limitations have managed the adoption of XAI, specifically SHAP, to optimize the model output interpretability and facilitate better insights on the way individual features form classification decision. Moreover, several previous studies were having a lack of rigorous, algorithm-driven feature selection stages, such as RFE [24]. This might turn the feature space into a burdened feature with irrelevant or overcycling variables, leading to feature redundancy, multi-collinearity, and the overfitting risk. These circumstances show the requirements to build a PCOS classification model which has high predictive performance completed with maximized feature selection and relevant clinic interpretation. The dataset applied consisted of 541 patient data, followed by 42 informative attributes, referring to hormonal, clinical, anthropometric, until lifestyle conditions. The informative attributes implemented here are depicted through Table 1.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 1. Informative attributes of the PCOS dataset

No	Variable	Unit or Category	Data Type	Description
1	'PCOS (Y/N)'	Yes / No	'Categorical'	PCOS status variable
2	'Age (yrs)'	Years	'Numerical'	Patient's age
3	'Weight (Kg)'	Kilogram	'Numerical'	Patient's weight
4	'Height (Cm)'	Centimeter	'Numerical'	Patient's height
5	'BMI'	kg / m²	'Numerical'	Patient's body mass index
6	'Blood Group'	A / B / AB / O	'Categorical'	Patient's blood group
7	'Pulse rate (bpm)'	bpm	'Numerical'	Patient's pulse rate
8	'RR (breaths/min)'	Times / minute	'Numerical'	Patient's respiratory rate
9	'Hb (g/dl)'	g/dl	'Numerical'	Hemoglobin level
10	'Cycle (R/I)'	Regular / Irregular	'Categorical'	Menstrual cycle pattern
11	'Cycle length (days)'	Days	'Numerical'	Menstrual cycle length
12	'Marriage Status (Yrs)'	Years	'Numerical'	Duration of marriage
13	'Pregnant (Y/N)'	Yes / No	'Categorical'	Pregnancy status
14	'No. of abortions'	Count	'Numerical'	History of abortions
15	'I β-hCG (mIU/mL)'	mIU / L	'Numerical'	β-hCG level (first stage)
16	'II β-hCG (mIU/mL)'	mIU / L	'Numerical'	β-hCG level (second stage)
17	'FSH (mIU/mL)'	mIU / L	'Numerical'	FSH level
18	'LH (mIU/mL)'	mIU / L	'Numerical'	LH level
19	'FSH/LH'	Ratio	'Numerical'	FSH to LH ratio
20	'Hip (inch)'	Inch	'Numerical'	Hip circumference
21	'Waist (inch)'	Inch	'Numerical'	Waist circumference
22	'Waist–Hip Ratio'	Rasio	'Numerical'	Waist-to-hip ratio
23	'TSH (mIU/L)'	mIU / L	'Numerical'	TSH level
24	'AMH (ng/mL)'	ng / mL	'Numerical'	AMH level
25	'PRL (ng/mL)'	ng / mL	'Numerical'	Prolactin hormone level
26	'Vit D3 (ng/mL)'	ng / mL	'Numerical'	Vitamin D3 level

No	Variable	Unit or Category	Data Type	Description
27	'PRG (ng/mL)'	ng / mL	'Numerical'	Progesterone hormone level
28	'RBS (mg/dl)'	mg / dl	'Numerical'	Random blood glucose level
29	'Weight gain (Y/N)'	Yes / No	'Categorical'	History of weight gain
30	'Hair growth (Y/N)'	Yes / No	'Categorical'	Excessive hair growth
31	'Skin darkening (Y/N)'	Yes / No	'Categorical'	Skin hyperpigmentation
32	'Hair loss (Y/N)'	Yes / No	'Categorical'	Hair loss
33	'Pimples (Y/N)'	Yes / No	'Categorical'	Acne
34	'Fast Food (Y/N)'	Yes / No	'Categorical'	Fast food consumption
35	'Reg. Exercise (Y/N)'	Yes / No	'Categorical'	Regular physical exercise
36	'BP Systolic (mmHg)'	mmHg	'Numerical'	Systolic blood pressure
37	'BP Diastolic (mmHg)'	mmHg	'Numerical'	Diastolic blood pressure
38	'Follicle No. (L)'	Count	'Numerical'	Number of left ovarian follicles
39	'Follicle No. (R)'	Count	'Numerical'	Number of right ovarian follicles
40	'Avg. F size (L) (mm)'	mm	'Numerical'	Average size of left ovarian follicles
41	'Avg. F size (R) (mm)'	mm	'Numerical'	Average size of right ovarian follicles
42	'Endometrium (mm)'	mm	'Numerical'	Endometrial thickness

This study was implemented in Python using Google Colaboratory. The key libraries used in this study are summarized in Table 2.

Table 2. Key libraries

Library	Version	Purpose
pandas	2.2.2	Data manipulation and preprocessing.
numpy	2.0.2	Numerical computation and array operations.
scikit-learn	1.6.1	Machine learning pipeline including train-test split, Random Forest classifier, RFE feature selection, and evaluation metrics.
Shap	0.51.0	Explainable AI via SHAP values for model interpretability.
Matplotlib	3.10.0	Visualization.
seaborn	0.13.2	Statistical data visualization including heatmap and distribution plots.

2.2. Preprocessing Data

The preprocessing stage began with loading the PCOS dataset from an Excel file using pd.read_excel(), resulting in 541 records and 44 columns. Column names were standardized using .str.strip() to remove extra whitespace, and unnamed columns resulting from Excel export artifacts were identified and removed using .drop(). All object-type columns were subsequently converted to numeric format using pd.to_numeric() with errors='coerce', converting non-numeric entries to NaN for subsequent imputation. Missing value inspection using .isna().sum() identified four columns with one missing entry each, namely Fast food (Y/N), AMH (ng/mL), II beta-HCG (mIU/mL), and Marriage Status (Yrs). Imputation was applied using .fillna(), filling all numeric columns with their respective median values; no separate mode-based imputation for categorical columns was performed within the preprocessing stage, as all remaining columns had already been converted to numeric format prior to this step. Zero total missing values after imputation were confirmed by .isnull().sum().sum(). Outlier detection was subsequently performed using the IQR method, where values below Q1 − 1.5×IQR or above Q3 + 1.5×IQR were capped to their respective boundary values using .clip() to limit the influence of extreme observations without discarding any records.

2.3. Modeling

2.3.1. Split Data

The dataset was divided into training and testing sets using an 80:20 ratio via scikit-learn's train_test_split(), a proportion widely adopted in machine learning studies as it provides sufficient training data for model generalization while retaining an adequate

portion for unbiased evaluation. Stratified sampling was implemented through the stratify=y parameter. The goal is to preserve the class distribution of the targetted variable across both subsets. Meanwhile, random_state=42 was managed to guarantee reproducibility. Every informed result is based on this single stratified 80:20 split with no repetitive cross-validation demonstrated. The stratified split was valued as appropriate for dataset size given evaluation. The data training set was applied for feature selection along with model fitting. Meanwhile, the test set was applied to exclusively evaluate the final state.

2.3.2. Feature Selection

Two feature selection approaches were implemented and compared here:

1) Correlation Filter (CF)

In the Correlation Filter (CF) feature selection step, the absolute correlation threshold among every numerical feature in the data training and the target variable 'PCOS (Y/N)' is measured with the help of the corr() function. Features with an absolute correlation above the 0.1 threshold are retained as the selected feature subset, while features below the threshold are eliminated from the subsequent process. The selected features will be used to form the training data subset X_train_cf and the testing data subset X_test_cf as inputs in the RF + CF model training stage.

2) Recursive Feature Elimination (RFE)

RFE was applied using a base Random Forest estimator configured with n_estimators=200, random_state=42, and class_weight='balanced'. The RFE procedure iteratively trains the base estimator and eliminates the least important feature at each step (step=1) until the target number of features is reached. The number of features to retain was set to 21, chosen to match the CF output for a fair and controlled comparison between the two selection strategies [24]. chosen to match the number of features produced by the CF approach, ensuring a controlled and fair comparison between the two selection strategies rather than representing the globally optimal feature count. The features retained by RFE were identified through the rfe.support_ mask and used to transform both the training X_train_rfe and testing X_test_rfe sets prior to model training. A smaller value of n_estimators=200 was used for the RFE base estimator to reduce computational overhead during the iterative feature elimination process, as RFE requires repeated model fitting at each elimination step.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

2.3.3. Training Model

Following feature selection, the final Random Forest classifier was trained independently for each feature selection pipeline (RF+CF and RF+RFE) using the configuration: n_estimators = 300, random_state = 42, class_weight = 'balanced', and n_jobs = -1. The class_weight='balanced' parameter was applied to address the class imbalance present in the dataset by automatically adjusting class weights inversely proportional to class frequencies. The n_estimators value of 300 was selected for the final model higher than the 200 used in the RFE base estimator to ensure greater ensemble diversity and more stable predictions during final classification, while remaining computationally feasible. Formal hyperparameter maximized procedure like grid or randomized search performed is not found; the informed configuration shows a deliberately chosen setting according to common practices and preliminary exploration. Other hyperparameters were managed to scikit-learn setting. The RF+CF model was trained by having a feature subset arranged by the Correlation Filter (X_train_cf, y_train). Meanwhile, the RF+RFE model was trained with a feature subset arranged by Recursive Feature Elimination (X_train_rfe, y_train).

2.3.4. Model Testing

Every trained model was evaluated by using .predict() to the respective held-out test subset, arranging predicted class labels which were subsequently used in comparison against the ground truth. Performance was quantified using four metrics accuracy, precision, recall, and F1-score computed via accuracy_score(), precision_score(), recall_score(), and f1_score(), with detailed per-class breakdowns reported through classification_report(). Prediction outcomes were then illustrated with a confusion matrix rendered through ConfusionMatrixDisplay, allowing inspection of true positive, true negative, false positive, and false negative classifications for every model variety.

2.3.5. Explainable AI – SHAP

Analysis of model interpretability was completed with the help of SHAP and shap.TreeExplainer [18]. This was separatedly initialized for the RF+CF model (explainer_cf) and the RF+RFE model (explainer_rfe). The SHAP analysis for the RF+RFE model was exclusively implemented to the last Random Forest classifier trained on the RFE-selected feature subset, making sure that the outputs of interpretation has reflected only the features retained after elimination. SHAP values were measured on the data test

of every model and transformed for the positive class (PCOS = 1) to identify every feature contribution to the predicted results.

The arranged illustrations include two kinds: (1) a summary plot, illustrating the magnitude and direction of every feature's contribution to individual predictions, and (2) a bar plot, presenting the mean of definite SHAP values as representation of the importance of global feature. Both of these illustrations were presented side by side to show the comparison of RF+CF and RF+RFE models' comparative interpretation in determining the most dominating clinical features for PCOS categorization.

3. RESULTS AND DISCUSSION

3.1. Model Training

Every model trained was reviewed through the held-out test set comprising 20% of the overall dataset, which includes 109 records. The presentation of classification between RF, RF+CF, and RF+RFE was computed by having four metrics, including accuracy, precision, recall, and F1 score. The results were summarized and shown through Table 3.

Table 3. Evaluation Result

Model	Accuracy	Precision	Recall	F1-Score
Rf (Baseline)	92.66%	96.67%	80.56%	87.88%
RF+CF	91.74%	93.55%	80.56%	86.57%
RF+RFE	92.66%	93.75%	83.33%	88.24%

The matrix of confusion for every model was arranged through ConfusionMatrixDisplay. It can be viewed through Figure 2, where a per-class breakdown indicates the prediction of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The final confusion matrix values for every model as summarized through Table 4.

Table 4. Confusion Matrix Values for All Models.

Model	TP	TN	FP	FN
Rf (Baseline)	29	72	1	7
RF+CF	29	71	2	7

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Model	TP	TN	FP	FN
RF+RFE	30	71	2	6

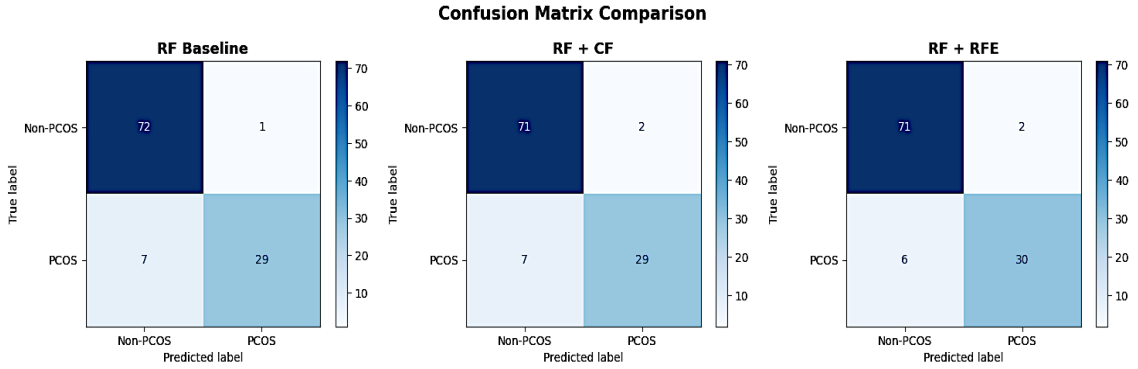

Figure 2. Model Confusion Matrix

3.2. Model Evaluation

As can be reviewed from Table 3, the RF+RFE model presented powerful overall performance related to recall and F1-score. In this case, recall is at the percentage of 83.33% and F1 is at 88.24% while distributing the similar accuracy of 92.66% as the RF baseline. To be compared, the RF+CF gained an accuracy of 91.74%, precision of 93.55%, recall of 80.56%, and F1-score of 86.57%. The relatively higher recall analyzed in RF+RFE is specifically significant within the clinical context of PCOS identification, as minimizing false negatives is crucial to avoid false diagnoses. In a screening-oriented clinical context, missed diagnoses of PCOS are typically characterized by false negatives, the consequences of which are far more serious than those of false positives. This is because undetected cases can progress to long-term metabolic complications, including type 2 diabetes, cardiovascular disease, and infertility if not treated in a timely manner. The improved detection achieved by RF+RFE responsibly provides a framework for clinical decision-making that prioritizes patient safety over an overly conservative statistical approach. It is crucial to note that the improvement of RF+RFE accuracy over the RF baseline cannot be said to be main finding. Instead, it is the main improvement for the recall and F1-score, showing more sensitive sense to identify true PCOS-positive cases. The gap of performance between those two models encourages RFE to produce a more discriminative feature subset, efficiently selecting noise while retaining predictors which are clinically relevant. The implementation of class_weight='balanced' for both models had contribution to mitigate the effects of class imbalance exist in the dataset, as

illustrated in the justified precision and recall scores across both PCOS-positive along with PCOS-negative categories [14].

RF baseline, trained on all 42 features, gained a 92.66% accuracy and the highest precision of 96.67%. On the other hand, this high precision arrived at the cost of a quite low recall at the percentage of 80.56%, showing that the baseline model kept being overly traditional in determining PCOS-positive issues despite integrating access to the overall set of features. The excessively high precision relative to recall encourages that clinically irrelevent features inclusivity introduced noise which biased the boundary of decision related to the dominating negative class, creating a higher rate of false negatives which is undesirable in the context of clinical screening.

The results of evaluation affirm that RF+RFE receives the most favorable diagnostic profile among the three models. As illustrated through Table 3, RF+RFE and RF Baseline distribute the equal accuracy of 92.66%, while the main benefit of RF+RFE can be seen through its increased recall (83.33% vs. 80.56%) and F1-score (88.24% vs. 87.88%) over the RF baseline, presenting that RFE-based feature selection developed model sensitivity with no overall sacrifice related to the accuracy of classification. Even though the RF Baseline received the highest precision (96.67%), this arrives at the cost of a higher false negative rate (FN=7), which means more factual PCOS issues were still not identified; in the context of clinical screening like trade-off is commonly unfavorable, as the result of a neglected diagnosis outweighs that of a false alarm, rendering recall and F1-score more appropriate indicators of model usage. RF+CF, while managing a competitive precision of 93.55%, yielded the lowest accuracy (91.74%) and F1-score (86.57%) among all varieties, encouraging that correlation-based filtering alone was not more efficient to isolate the most discriminative feature subset for PCOS categorization. The findings of this study have clinical significance in that correlation-based screening, although computationally relatively simple, treats each feature and the target variable separately and linearly; this causes the method to overlook features whose diagnostic value emerges through interactions with other variables, which constitutes a limitation. Specifically in the case of PCOS, hormonal dysregulation involves complex multivariate interactions along the hypothalamic-pituitary-ovarian axis, and the method is inherently less capable of capturing these interactions, placing it at a disadvantage. It is crucial to recognize that the recall increase of RF+RFE over the RF baseline along with RF+CF relates to a decrease

of only one false negative (from FN=7 to FN=6) in the held-out test set of 109 records. While this is interpreted into a meaningful decrease in the false negative rate between 19.4% to 16.7%, the statistical strength of this increase is limited by the small test set size, and wider validity on larger or multi-center datasets is suggested before arranging definitive clinical summaries. Overall, the findings present that feature selection takes such a crucial and meaningful part in model performance, with RFE offering a more refined feature representation which encourages the model's potential to appropriately determine PCOS-positive issues. These also encourage the adaptation of RF+RFE as the chosen classification model in this study, especially given the clinical priority of optimizing recall within the PCOS screening scenarios.

3.3. Explainable AI – Shap

The analysis of SHAP was implemented exclusively to the last trained RF+CF and RF+RFE classifiers on their respective held-out test subsets, facilitating two-level interpretation: at the global level, determining which features are most affecting among the entire test set, and at the local level, presenting the direction and magnitude of every feature's contribution to individual patient predictions. The findings show that 'Follicle No. (R)' along with 'Follicle No. (L)' hold the highest mean absolute SHAP values towards both models, classifying them as the most affective predictors in PCOS classification [14]. An elevated follicle count exerts a robust positive effect on PCOS predictions, stable with the Rotterdam diagnostic criteria that place considerable diagnostic weight on polycystic ovarian morphology [8]. The results of machine learning-derived feature importance and established clinical criteria show alignment between the established levels of feature importance. This is highly valuable because it provides face validity for the model's decision logic, indicating that RF+RFE is not merely a statistical artifact but has also learned clinically coherent patterns. In a translational context, this means clinicians may, in principle, be able to trace these predictions back to features they recognize as diagnostically important, which is a prerequisite for building trust in AI-powered screening tools. Clinical symptoms integrated with hyperandrogenism, containing 'Weight gain (Y/N)', 'Skin darkening (Y/N)', and 'Hair growth (Y/N)', also present significant positive contributions towards both models, projecting their developed part in the PCOS clinical presentation

A main interpretive differentiation between the two models emerges from the hormonal contributors. In the RF+CF model, hormonal indicators like FSH, LH, and FSH/LH ratio are not present from the SHAP analysis since they were excluded during the CF feature selection stage. Meanwhile, the RF+RFE model incorporates 'AMH (ng/mL)' and 'FSH/LH ratio' as meaningful SHAP contributors, projecting the model's ability to acknowledge clinically significant endocrine disruptions that CF was unable to capture systematically. This distinction emerges the argument that RFE creates a more hormonally finished and physiologically grounded feature set, better in line with the acknowleged PCOS pathophysiology [25].

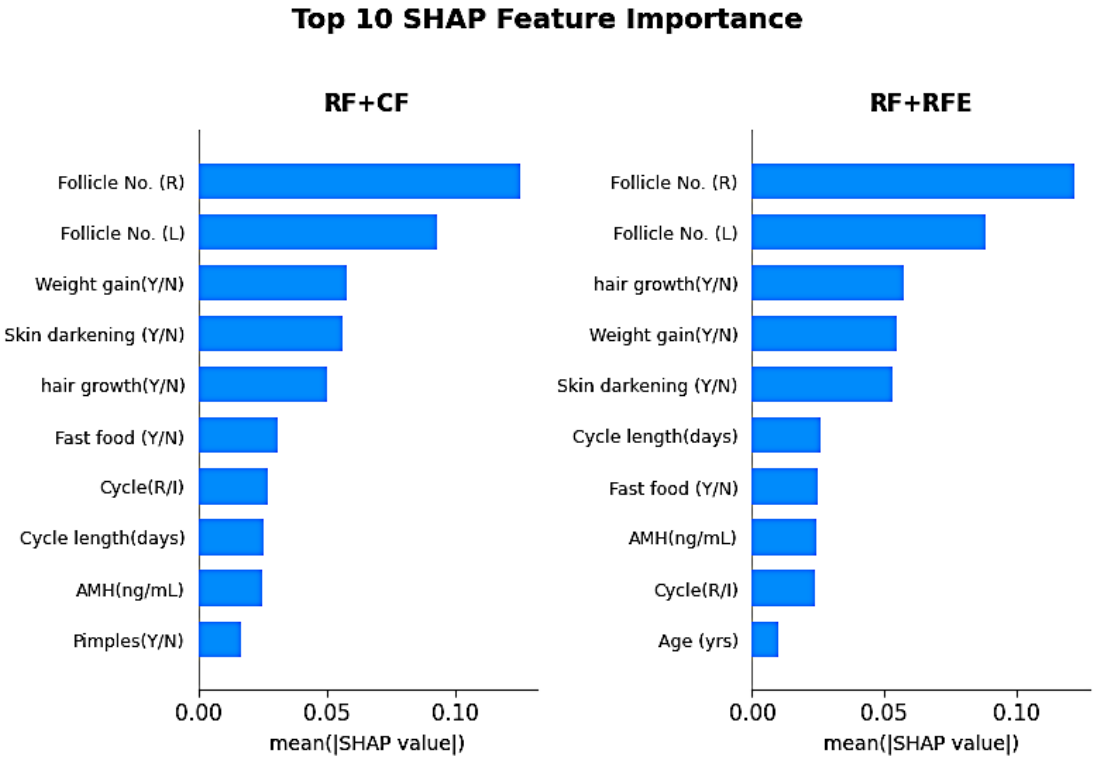

Figure 3. Top 10 Feature Importance

Figure 3 presents the ten most affective predictors as identified by their mean absolute SHAP values for both models. As presented through Figure 3, the majority of follicular count variables is stable across both models, affirming their central diagnostic role. In addition, the RF+RFE bar plot underlines 'AMH (ng/mL)', 'Cycle length (days)', and 'Age (yrs)' across the top contributors — variables absent from the CF SHAP ranking — while 'Pimples (Y/N)' appears exclusively in the CF ranking, presenting that the RFE-selected feature set depicts a wider and more clinically representative spectrum of PCOS indicators outside the morphological results.

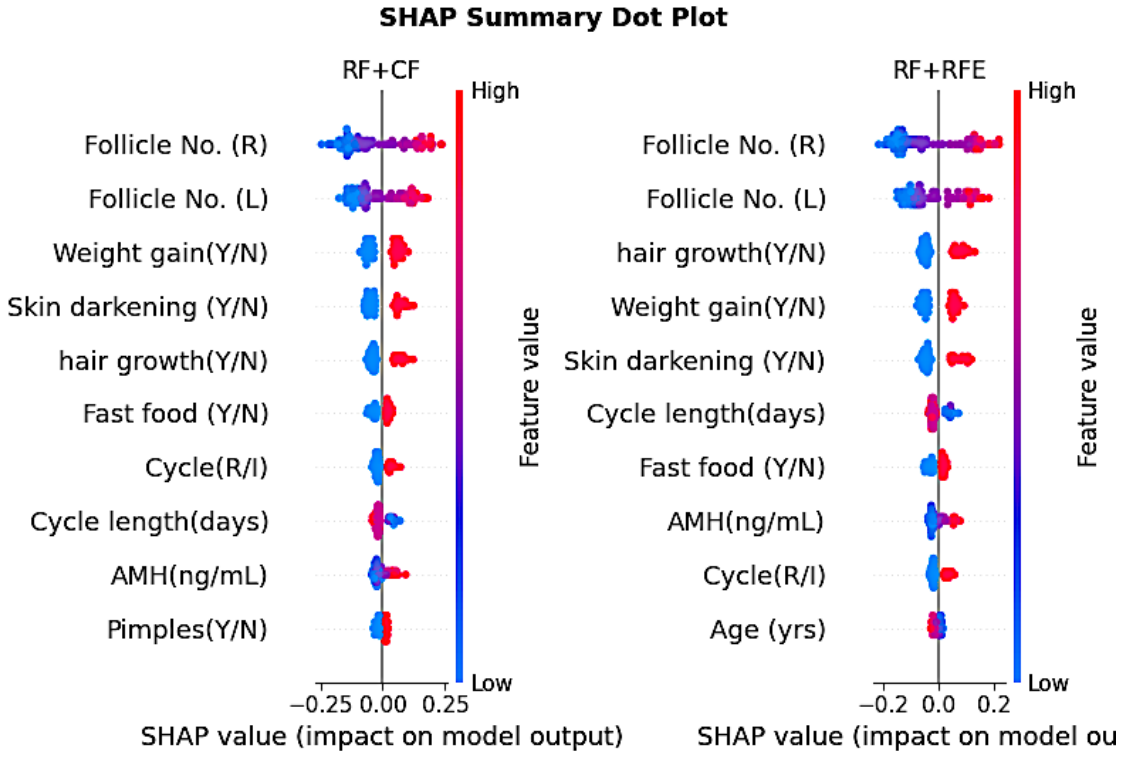

Figure 4. SHAP Dot Plot

Figure 4 demonstrates the SHAP summary dot plots for both models, presenting the distribution, direction, and magnitude of every feature's contribution to individual predictions. As depicted from Figure 4, high values of 'Follicle No. (R)' along with 'Follicle No. (L)', recaptured by red dots on the positive SHAP axis. In contrast, high values of 'Cycle (R/I)' push predictions toward the negative class, showing the clinical observation that regular menstrual sequence might decrease the PCOS likelihood. In the RF+RFE summary plot, hormonal features like 'AMH (ng/mL)' along with 'FSH/LH' exhibit more understandable distributional patterns and broder SHAP value length, in comparison to the CF model, showing that those endocrine markers contribute more variously to individual-level predictions, encouraging the interpretation that RF+RFE operates with a more physiologically grounded decision logic suitable for transparent clinical decision support.

Figure 5 presents the grouped bar chart comparing the four evaluation metrics across all three models. As illustrated in Figure 5, the performance advantage of RF+RFE is most evident in the Recall and F1-score bars, while the notably higher Precision bar of Pure RF relative to its Recall bar visually confirms its conservative prediction tendency. The bar

heights of RF+RFE across recall and F1-score indicate a more clinically appropriate classifier compared to both the baseline and RF+CF.

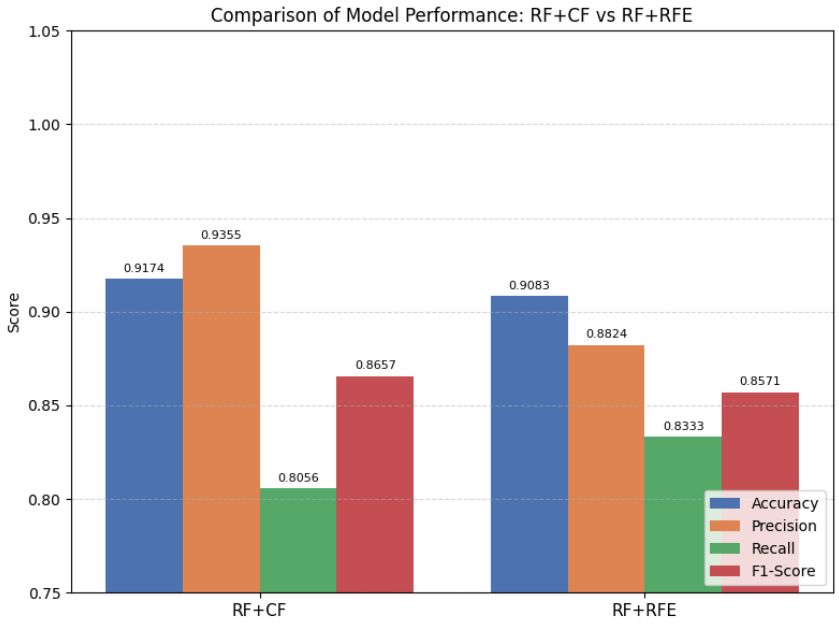

Figure 5. Comparation Model

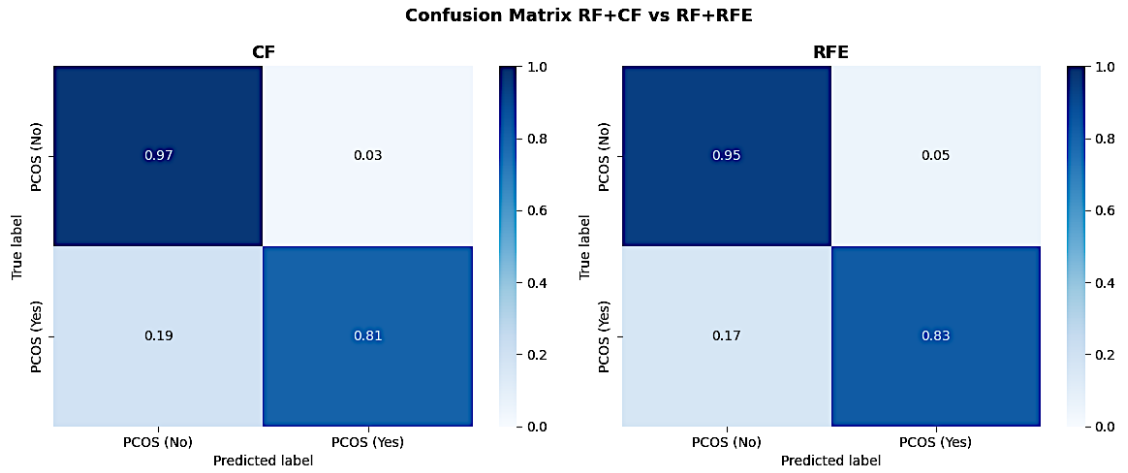

Figure 6. Confusion Matrix

Figure 6 presents the confusion matrices for RF+CF and RF+RFE displayed side-by-side. As illustrated in Figure 6, the RF+CF matrix shows 7 false negatives out of 36 positive cases, corresponding to a false negative rate of 19.4%, meaning nearly one in five PCOS-positive patients would be missed in a screening scenario. The RF+RFE matrix reduces this to 6 false negatives, lowering the false negative rate to 16.7%. This reduction

represents a single case in the held-out test set of 109 records, and while it reflects a clinically meaningful direction of improvement in screening sensitivity, its statistical significance is limited by the small test set size and should be interpreted with appropriate caution.

3.4. Discussions

The experimental results demonstrate that both feature selection approaches effectively reduced the feature space from 42 informative attributes to 21 features while maintaining high classification performance [14]. In general, the RF+RfE model, which achieved an accuracy of 92.66%, is consistent with other Random Forest-based PCOS classification studies. Tiwari et al. [5] reported a comparable accuracy of 92.4% using Random Forest at a correlation threshold of 0.8, while Nasim et al. [10] achieved an accuracy of 93.79% using Random Forest in a broader multi-algorithm comparison on a similar PCOS clinical dataset. However, previous studies did not employ systematic algorithmic feature selection such as RFE, nor SHAP-based explainability; while direct comparison is limited by differences in datasets and experimental setups, these two aspects suggest that comparable or superior recall may potentially be achieved alongside transparent feature-level explanations, subject to further validation across diverse datasets. The RF baseline, operating on all 42 features, achieved an accuracy of 92.66% but exhibited a markedly imbalanced precision-recall profile (96.67% precision vs. 80.56% recall), suggesting that the inclusion of clinically irrelevant features introduced noise that biased the model toward over-predicting the negative class. This finding highlights a critical limitation of using unselected full-feature models in clinical screening contexts, where high precision at the expense of recall is clinically unacceptable.

The RF+RFE model dealed with this limitation by accepting the equal accuracy of 92.66% while increasing recall from 80.56% to 83.33% and F1-score from 87.88% to 88.24%, in comparison to the RF baseline, showing the primary performance gain of the suggested approach rather than an accuracy increase. This re-balancing was gained through the 50% decrease in feature dimensionality, which erased clinically irrelevant variables which put some contribution noise to the baseline model's boundary of decision. The powerful performance of RF+RFE over RF+CF might be attributed to two complementary factors. First, RFE's iterative elimination sequel led by Random Forest feature urgency scores is able to determine features which put some contribution through non-linear and

interaction-based pathways which CF's linear correlation criterion inherently cannot detect. Furthermore, the hormonal variables retained by RFE, specifically FSH, LH, FSH/LH ratio, and AMH, are mechanistically implemented in PCOS pathogenesis through hypothalamic-pituitary-ovarian axis dysregulation [5], and their inclusivity created a decision boundary which better projects the endocrine condition complexity.

From a clinical perception, recall is the most crucial metric in PCOS screening, as false negatives show neglected diagnoses which may hinder treatment and improve the risks of long-term complications, encompassing type 2 diabetes, cardiovascular disease, until infertility [1],[5]. The RF+RFE model received the highest recall of 83.33% among all three models, decreasing false negatives from 7 to 6 compared to both RF and RF+CF. On the other hand, it should be recognized that this improvement has a relation to a distinction of only one false negative in a test set of 109 records, that limits the statistical strength of this result. Thus, the observed recall improvement should be projected as a promising directional finding rather than a conclusive clinical claim, and validation on bigger and more various datasets is highly suggested. Limitations are not unique to this study; Arora et al. [22] noted in their systematic review of machine learning diagnostic methods for PCOS that most models were trained and validated using datasets from a single research center, raising consistent concerns regarding generalizability to populations with varying diagnostic criteria and demographic profiles. Further research that extends the RF+RFE+SHAP pipeline to multi-center datasets would allow for a more rigorous assessment of the stability of the identified feature selection rankings—particularly for hormonal variables—across ethnically and clinically diverse cohorts.

The SHAP analysis affirmed the majority of follicular variables in PCOS prediction and facilitated transparent global and local interpretations for both models. The consistency of 'Follicle No. (R)' and 'Follicle No. (L)' as the top SHAP contributors among both models affirms the alignment of the machine learning pipeline with developed clinical diagnostic criteria, lending medical credibility to the logic of model's decision The findings of this study are consistent with those of Elmannai et al. [1], who identified follicular morphology as the primary predictor in their XAI-based PCOS detection model using optimized feature selection, reinforcing that polycystic ovarian morphology carries the strongest discriminative signal across various dataset configurations. The prominence of features related to hyperandrogenism—such as weight gain, hair growth, and skin darkening—in

the SHAP rankings of both models reinforces the findings of Sreejith et al. [11] and Alagarsamy et al. [4] who demonstrated that clinical symptom-based variables consistently rank at the top as contributors in ensemble-based PCOS classifiers. The additional contributions of AMH and the FSH/LH ratio in the RF+RFE model, which are absent in the RF+CF model, align with the endocrinological evidence cited by Teede et al. [4], who identified these hormonal markers as central to the pathophysiology of PCOS, suggesting that the RFE feature set may potentially be more physiologically aligned with established PCOS pathophysiology, though this interpretation remains subject to further validation on larger and more diverse datasets. The additional SHAP contributions of 'AMH (ng/mL)' and 'FSH/LH ratio' in the RF+RFE model then encourage the clinical completeness of its feature set, straightforwardly facing the interpretability limitation of black-box ensemble models and reinforcing the suitability of the suggested framework for transparent clinical decision support 24-26 [18].These results are in line with previous studies informing RF efficiency in PCOS classification [1],[3],[6],[13], while extending the evidence base by presenting that systematic RFE-based feature selection creates a more clinically equal and interpretable model than unselected full-feature approaches [24].

4. CONCLUSION

This study developed a PCOS classification model using the Random Forest algorithm combined with two feature selection methods, Correlation Filter (CF) and Recursive Feature Elimination (RFE), evaluated on a clinical dataset of 541 patients with 42 informative attributes. Both methods successfully reduced the feature space to 21 variables. The RF baseline trained on all 42 features achieved the same accuracy of 92.66% as RF+RFE but exhibited a markedly imbalanced precision-recall profile (96.67% precision vs. 80.56% recall), indicating an overly conservative tendency in identifying true PCOS-positive cases. The RF+RFE model achieved the same accuracy as the RF baseline (92.66%) while improving recall from 80.56% to 83.33% and F1-score from 87.88% to 88.24%, with these recall and F1-score gains representing the primary performance contribution of the proposed approach, outperforming the RF+CF model which obtained 91.74%, 93.55%, 80.56%, and 86.57% respectively. 'Follicle No. (L)', consistently ranked as the most influential predictors across both models, followed by hyperandrogenism-related symptoms such as weight gain, skin darkening, and hair growth. The primary

contribution of this study lies in the systematic integration of RFE-based feature selection and SHAP-based explainability into a unified Random Forest pipeline, which improved recall and F1-score performance and enhanced the clinical interpretability of PCOS classification outcomes. These findings support the potential of combining RF with RFE and interpretable AI techniques as a screening-support approach based on structured clinical data. However, this model has not been validated as a clinical diagnostic tool, and these results are derived from a single dataset of 541 patients broader validation on larger and more diverse clinical populations is still required. Other research in the future is suggested to validate the proposed framework on multi-center datasets, integrate additional clinical modalities, and explore integration with clinical decision support systems to strengthen its applicability in real-world healthcare fields.

REFERENCES

[1] H. Elmannai *et al.*, "Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence," *Diagnostics*, vol. 13, no. 8, pp. 1–21, 2023, doi: 10.3390/diagnostics13081506.

[2] S. Arora, Vedpal, and N. Chauhan, "*Polycystic Ovary Syndrome (PCOS) diagnostic methods in machine learning: a systematic literature review*", vol. 84, no. 16. Springer US, 2025. doi: 10.1007/s11042-024-19707-6.

[3] S. Ahmed *et al.*, "A Review on the Detection Techniques of Polycystic Ovary Syndrome Using Machine Learning," *IEEE Access*, vol. 11, pp. 86522–86543, 2023, doi: 10.1109/ACCESS.2023.3304536.

[4] M. Alagarsamy, N. Shanmugam, D. P. Mani, M. Thayumanavan, K. K. Sundari, and K. Suriyan, "Detection of Polycystic Syndrome in Ovary Using Machine Learning Algorithm," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 1, pp. 246–253, 2023.

[5] S. Tiwari *et al.*, "SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning," *Expert Syst. Appl.*, vol. 203, no. May, 2022, doi: 10.1016/j.eswa.2022.117592.

[6] J. Lim *et al.*, "Machine learning classification of polycystic ovary syndrome based on radial pulse wave analysis," *BMC Complement. Med. Ther.*, vol. 23, no. 1, pp. 1–15, 2023, doi: 10.1186/s12906-023-04249-5.

[7] C. Aulia *et al.*, "Analisis Pola Gejala Pcos Menggunakan Algoritma K-Means Clustering," *JOISIE (Journal Inf. Syst. Informatics Eng.*, vol. 9, no. 1, pp. 91–99, 2025,

[Online]. Available: https://www.ejournal.pelitaindonesia.ac.id/ojs32/index.php/JOISIE/article/view/4939

[8] H. J. Teede *et al.*, "Recommendations From the 2023 International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary Syndrome," *J. Clin. Endocrinol. Metab.*, vol. 108, no. 10, pp. 2447–2469, 2023, doi: 10.1210/clinem/dgad463.

[9] H. Yang *et al.*, "Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators," *Inf. Fusion*, vol. 75, no. February, pp. 140–149, 2021, doi: 10.1016/j.inffus.2021.02.015.

[10] S. Nasim, M. S. Almutairi, K. Munir, A. Raza, and F. Younas, "A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics," *IEEE Access*, vol. 10, no. September, pp. 97610–97624, 2022, doi: 10.1109/ACCESS.2022.3205587.

[11] S. Sreejith, H. Khanna Nehemiah, and A. Kannan, "A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier," *Healthc. Anal.*, vol. 2, no. March, p. 100102, 2022, doi: 10.1016/j.health.2022.100102.

[12] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "The accuracy of Random Forest performance can be improved by conducting a feature selection with a balancing strategy," *PeerJ Comput. Sci.*, vol. 8, pp. 1–15, 2022, doi: 10.7717/PEERJ-CS.1041.

[13] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "Feature selection to increase the random forest method performance on high dimensional data," *Int. J. Adv. Intell. Informatics*, vol. 6, no. 3, pp. 303–312, 2020, doi: 10.26555/ijain.v6i3.471.

[14] S. Alam Suha and M. N. Islam, "Exploring the dominant features and data-driven detection of polycystic ovary syndrome through modified stacking ensemble machine learning technique," *Heliyon*, vol. 9, no. 3, p. e14518, 2023, doi: 10.1016/j.heliyon.2023.e14518.

[15] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *Int. J. Data Sci. Anal.*, vol. 20, no. 2, pp. 197–211, 2025, doi: 10.1007/s41060-024-00509-w.

[16] S. Ratnasingam and J. Muñoz-Lopez, "Distance Correlation-Based Feature Selection in Random Forest," *Entropy*, vol. 25, no. 9, 2023, doi: 10.3390/e25091250.

[17] M. Mohamad, A. Selamat, O. Krejcar, R. G. Crespo, E. Herrera-Viedma, and H. Fujita, "Enhancing big data feature selection using a hybrid correlation-based feature

selection," *Electron.*, vol. 10, no. 23, pp. 1–24, 2021, doi: 10.3390/electronics10232984.

[18] N. G. Rezk, S. Alshathri, A. Sayed, E. El-Din Hemdan, and H. El-Behery, "XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach," *Bioengineering*, vol. 11, no. 10, 2024, doi: 10.3390/bioengineering11101016.

[19] P. K. Mohanty, S. A. J. Francis, R. K. Barik, D. S. Roy, and M. J. Saikia, "Leveraging Shapley Additive Explanations for Feature Selection in Ensemble Models for Diabetes Prediction," *Bioengineering*, vol. 11, no. 12, pp. 1–19, 2024, doi: 10.3390/bioengineering11121215.

[20] O. O. Bifarin, "Interpretable machine learning with treebased shapley additive explanations: Application to metabolomics datasets for binary classification," *PLoS One*, vol. 18, no. 5 May, 2023, doi: 10.1371/journal.pone.0284315.

[21] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare," *AI*, vol. 4, no. 3, pp. 652–666, 2023, doi: 10.3390/ai4030034.

[22] T. Patil and S. Arora, "Survey of Explainable AI Techniques: A Case Study of Healthcare," *Lect. Notes Networks Syst.*, vol. 765 LNNS, pp. 335–346, 2023, doi: 10.1007/978-981-99-5652-4_30.

[23] D. Saraswat *et al.*, "Explainable AI for Healthcare 5.0: Opportunities and Challenges," *IEEE Access*, vol. 10, no. July, pp. 84486–84517, 2022, doi: 10.1109/ACCESS.2022.3197671.

[24] S. Xia and Y. Yang, "A Model-Free Feature Selection Technique of Feature Screening and Random Forest-Based Recursive Feature Elimination," *Int. J. Intell. Syst.*, vol. 2023, 2023, doi: 10.1155/2023/2400194.

[25] U. M. G and U. M. P, "SmartScanPCOS: A feature-driven approach to cutting-edge prediction of Polycystic Ovary Syndrome using Machine Learning and Explainable Artificial Intelligence," *Heliyon*, vol. 10, no. 20, 2024, doi: 10.1016/j.heliyon.2024.e39205.