Neural Computing and Applications
https://doi.org/10.1007/s00521-024-10171-9

ORIGINAL ARTICLE

PODBoost: an explainable AI model for polycystic ovarian syndrome detection using grey wolf-based feature selection approach

Poonam Moral[1] · Debjani Mustafi[1] · Sudip Kumar Sahana[1]

Received: 18 December 2023 / Accepted: 1 July 2024

Abstract
Polycystic Ovary Syndrome (PCOS) is a recurring endocrine disorder that primarily affects women of reproductive age. It is difficult to diagnose due to its heterogeneous characteristics and overlapping symptoms with other illnesses. As a result, accurate and trustworthy prediction models are required to detect *PCOS* early. This research work aims to develop ML methods that predict the risk of *PCOS* among women based on demographic and clinical features. The entire framework is divided into four phases: in Phase I of the study, the *SMOTE-Tomek Links (SMTL)* technique balances the data set by combining oversampling and undersampling approaches. A novel meta-heuristic-based feature selection approach, the *Grey Wolf Optimization (GWO)* method, has been employed to select the most crucial features from the dataset, explained in Phase II. Subsequently, in Phase III, a hybridized classifier *PODBoost* algorithm (Polycystic Ovarian Disorder Boosting algorithm) is devised for faithful early prediction of *PCOS* using the concepts of different classical supervised learning algorithms. Finally, *Explainable AI (XAI)* such as the *Local Interpretable Model-Agnostic Explanations (LIME)* tool has been implemented to interpret relevant predictions made by the proposed classifier. The proposed algorithm is examined utilizing numerous metrics such as *Accuracy, Error-Rate, ROC-AUC Score, Recall, Precision,* and *F1-Score*. Among all the evaluated models, the proposed hybridized model has shown an impressive performance with an exceptional accuracy of 97.42%, indicating its superiority by delivering outstanding results. Based on the findings, the novel meta-heuristic-based feature selection method significantly impacts the outcomes of the proposed hybridized *PODBoost* algorithm. This algorithm may be recommended to predict *PCOS* or other relevant diseases having datasets which are multimodal in nature.

Keywords Machine learning · PCOS · Boosting · Accuracy · Grey wolf optimization

1 Introduction

Polycystic Ovarian Disorder (POD) is a common disease caused due to the sex hormone imbalance [38, 39]. Elevated levels of the male hormone androgen in females lead to the formation of ovarian cysts. These lumps gradually emerge and impede the normal ovulation process. The obstruction in the ovulation process among women increases the probability of infertility. Approximately *5* to 10% of women are suffering from this disease [1, 48]. Different types of symptoms such as abnormal weight gain, migraine, irregular menstrual cycles, loss of hair, skin disease, etc., can be observed among adult females. *PCOS* disease is typically hereditary and is a life-threatening emergency [38]. Although geography and genes are the primary factors contributing to this disorder; an improper diet and infectious diseases can exacerbate the problem. The majority of women are unaware of the condition until they have a pregnancy test, contributing to a delayed diagnosis that frequently worsens the severity of the disease [4]. During pregnancy, women with *PCOS* face a risk of miscarriage that is over three times higher than that of

✉ Poonam Moral
phdcs10051.21@bitmesra.ac.in

Debjani Mustafi
debjani.mustafi@bitmesra.ac.in

Sudip Kumar Sahana
sudipsahana@bitmesra.ac.in

[1] Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, Jharkhand 835215, India

Published online: 30 July 2024

women without *PCOS*. Diagnosing *PCOS* does not involve a specific medical test. Instead, ultrasonography can reveal cysts, blood testing can assess androgen levels, and physical examination can analyze body hair. The prognosis of*PCOS* relies largely on observable physical symptoms and clinical testing, which inform the treatment process.

The development of numerous innovations in the area of biochemistry and healthcare technology has resulted in collection of massive amounts of data in the medical domain [12]. Thus, it has become a challenging task to identify sensible patterns from this voluminous data. Artificial Intelligence *(AI)* based approaches contribute a potential role in the healthcare domain [48]. The *ML* approach which is a subfield of *AI* works adequately well with voluminous, obscure datasets. Various ML algorithms may be implemented to investigate and analyze the data, convert it into an operational format for the medical process, and aid in determining the nature of numerous diseases. In the area of medicine, three types of data such as medical images, handwritten medical documents and genetics data are generally used to diagnose different diseases. Image processing, natural language processing, and statistical approaches are commonly employed to extract information from such data. Different supervised machine learning techniques such as classification, prediction, etc., have been employed to detect several diseases like diabetes, heart disease, breast cancer, autism and many more. Implementation of *ML* algorithms assist the medical practitioner to achieve satisfactory results which in turn help them to early prediction of a disease to reduce the severity and save several lives [53].

Despite the fact that various *ML-based* solutions have already been published in numerous research articles, progress in *PCOS* prognosis in recent years has been slow due to a scarcity of effectual and robust models. The most common machine learning approaches in medical diagnosis are classification systems which can discriminate between distinct groups. However, healthcare data like clinical data, medical images, etc., are comparatively higher in terms of dimensionality which might affect the classification accuracy. Moreover, high-dimensional data makes the model computationally intensive. It is very important to use multifaceted data for the development of a robust classification model to avoid overfitting phenomenon [12]. Traditional diagnosis is complicated and time-consuming due to the presence of several features that cause *PCOS* among women along with the difficulties in analyzing ultrasonography images [32].

1.1 Research contribution

The primary research contributions presented in this work include:

1. *SMOTE and Tomek Links (SMTL)* [50] technique is applied to reduce data imbalance, aiding in lowering the likelihood of overfitting. This has improved the accuracy of the minority class, which frequently results from random oversampling.
2. A population-based meta-heuristic *Grey Wolf Optimization* algorithm [35] is imposed to extract high-valued features, reducing the dataset from high-dimensional space to lower-dimensional space. Incorporating heuristic-based feature selection for early prediction of *PCOS* attains the state-of-the-art of the proposed classifier, minimizing the computational cost of the entire process.
3. *Exploratory Data Analysis (EDA)* [34] is accomplished to explore the insight of the data and to identify potential associations between features involved in predicting *PCOS* on the preprocessed dataset. Prominent features accountable for early prediction are visually represented to gain insight into the data.
4. The proposed hybrid classifier model is developed using the latent power of *Bagging and Boosting* techniques to improve the efficiency of existing ML algorithms.
5. *Explainable AI tool* such as *LIME* [14] is incorporated to understand the working of the proposed algorithm and its characteristics to enhance the predictive performance of the disease.

1.2 Organization of the study

Section 2 of the study systematically explores the relevant literature in detail. Afterward, Sect. 3 discusses the research methodology employed in the study along with the meta-heuristic-based feature engineering, which is one of the keystones of the investigation. Subsequently, in Sect. 4, the machine learning-based proposed model, in addition to some traditional classifiers, is investigated for the prediction of *PCOS*. Experimental results are analyzed, and evaluation measures are demonstrated extensively in Sect. 5. Finally, Sect. 6 summarizes the article by outlining the significant aspects of the proposed model and the future work directions.

2 Literature review

Artificial Intelligence and *Digital Healthcare* have made great strides in improving and enhancing medical diagnosis and treatment. This section examines the associated literature to our proposed work. Previous studies emphasize the importance of accurate diagnosis of *Polycystic Ovarian Disorder (POD)*. To review the existing work, this section

provides more emphasis on three perspectives: feature selection techniques, *PCOS* prediction models, and performance evaluation by *XAI*.

Bhardwaj et al. [7] highlighted *PCOS* prediction models using *ML* approaches to assist in self-diagnosis. To choose the most suitable attributes from the *PCOS* dataset (collected from the *Kaggle* dataset), the authors utilized statistical approaches such as Pearson correlation. A range of classification algorithms like *RF, MLP, XGB*, and SVM are used to enhance the model and achieved 93% accuracy with the *SVM* classifier. With the same instances, *Zigarelli et al.* [53] adopted the Categorical Boosting (CATB) algorithm for classification, securing 90.1% accuracy. *Tiwari S. et al.* [48] used non-invasive screening parameters to diagnose *PCOS* using clinical records provided by Kottarathil from the *Kaggle* repository. The experiment identified meaningful and distinguishing attributes based on the correlation coefficient, subsequently used to predict *PCOS*. The RF method was implemented for the prediction of *PCOS* with an accuracy of 93.25%. The effectiveness of the RF algorithm was assessed by employing an 'out-of-bag' error, and optimal values were chosen for tuning the parameters. *Zhang et al.* [52] conducted a systematic investigation into the correlation between the genetic characteristics of women and *PCOS*, with 233 patients with *PCOS* participating in the study. ML classifiers like *DT, KNN*, and *SVM* were used to identify new genes associated with *PCOS*. Among all investigated classifiers, the SVM outperformed with an accuracy rate of 80%. By performing data amalgamation and using feature selection methods such as the chi-square test and feature ranking, Aggarwal et al. [1] suggested a method for early prediction of *PCOS* disease. Two different datasets, such as the diabetes dataset and heart disease dataset, were fused to build a new dataset, which went through the feature extraction phase to get a reduced set of parameters (eight in number) and 985 records with the goal of identifying *PCOS*. Supervised and unsupervised learning models were utilized to explore the association between *PCOS* and other medical conditions, and the proposed work focused on identifying only important features for *PCOS* prediction. The study found that PCOS-affected women had higher chances of obesity, diabetes, and heart disease. *Hussain et al.* [25] introduced an electric theft detection method based on *Supervised Machine Learning (SML)* technique employing the CATB classifier in combination with the *Smote-Tomek Links (SMTL)* technique. Missing values in the obtained dataset were identified and assigned by the K-Nearest Neighbor technique, providing a realistic and accurate computation of the missing data. To manage the majority data class bias, the *SMTL* method was employed, which balanced the oversampling and undersampling approaches, and the Feature Extraction and Scalable Hypothesis techniques were employed to extract the most crucial attributes from the dataset. With the CATB algorithm, two categories, genuine and theft, were classified, and finally, the outcomes were interpreted by the tree-*SHAP* algorithm. In a research work conducted by *Danaei Mehr et al.* [12], traditional and ensemble models were used to analyze the *Kaggle PCOS* dataset to predict PCOS. The performance of several algorithms was tested with all attributes and reduced subsets of attributes formed by embedded, filter, and wrapper feature selection approaches. The findings exhibited that the feature selection approaches positively enhanced the performance of all algorithms. At last, it was observed that the Ensemble RF model, with reduced attributes through an embedded feature selection approach, outperformed other models with 98.89% accuracy.

The fundamental objective of the work presented by Kamel SR et al. [29] was to detect breast cancer by incorporating *SVM* classifier and *GWO-based* feature selection approach. The meta-heuristic-based *GWO* was used to extract the most efficient attributes to increase the efficiency of the breast cancer prediction technique. The hybrid approach witnessed the most effective measures in selecting relevant features by securing 100% accuracy, sensitivity, and specificity, surpassing other algorithms. Similarly, Al-Tashi et al. [3] suggested a method to determine the optimal attributes subset for determining cardiovascular disease using *GWO*. The suggested algorithm was executed in two stages: in the first stage, *GWO*was applied to select the best attributes for the identification of Coronary artery and in the second stage, the objective function of *GWO* was estimated by *SVM* classifier. This approach achieved 89.83% accuracy, 91% specificity, and 93% sensitivity rates, outshining the existing methods.

As *Explainable AI (XAI)* is increasingly seen as a means to ensure trustworthiness in healthcare by providing explanations for machine learning predictions. *Duell J et al.* [14] suggested *XAI* methods such as Scoped Rules, *LIME*, and *SHAP* to compute feature importance for ML predictions. *XAI* approaches improve the comprehension, interpretability, and dependability of model predictions. In this study, the authors evaluated the performance of these *XAI* methods on a large-scale electronic health record dataset with the goal of understanding the fatality rate for lung cancer. The comparison findings showed that the *XAI* models could generate insightful feature importance, emphasizing the need for domain experts to assess the reliability of the *XAI* models. As presented by *Elmannai et al.* [16], explainability is categorized into two levels: global explainability and local explainability. The final conclusion is explained across all data points via global explainability. It offers a cursory investigation of global fidelity. In terms of all samples, local loyalties could

explain. More precise explanations regarding the model could be obtained. The authors implemented the concept of Stacking ML models to predict *PCOS*, and the model was optimized by the Bayesian optimizer approach. The risk factors associated with polycystic ovarian disorder are estimated by cicek et al. [11] using the Random Forest model, which was further explained by one of the popular XAI tools, the LIME method. The results obtained by the method were examined, and it was observed that the values of Follicle R. and Follicle L. affect the prediction of PCOS. Exploratory Data Analysis is an important step in comprehending complex datasets, comprising interactive exploration through filtering, aggregation, and visualization. Nasim S. et. al. [38] combined EDA with advanced machine learning to detect PCOS, highlighting Gaussian Naive Bayes as a standout. Automation of EDA, driven by machine learning models, was discussed by Ganie SM et al. [20]. The researchers introduced an Ensemble Learning (EL) based approach for early Type-II diabetes prediction via lifestyle indicators, highlighting the significance of EDA to improve the quality of data. Exceptional results 99.4% were achieved with Bagged Decision Trees in healthcare applications.

However, further investigation is necessary to assess the efficacy of various machine learning classifiers on many datasets and to explore the possibility of unifying multiple *ML* models to boost the accuracy of *PCOS* prediction. This proposed method aims to construct a reliable algorithm for precisely forecasting *PCOS* in women based on medical records or clinical information.

3 Research methodology

The suggested *PCOS* detection framework is demonstrated in this section. The entire framework is divided into five primary phases: data pre-processing, hyperparameter tuning, model generation, model evaluation, and model elucidation. The entire architecture of the proposed study is exhibited in Fig. 1, and the illustration of each step is subsequently portrayed in the subsequent subsections.

3.1 PCOS dataset

This study makes use of the *PCOS* dataset available in the *Kaggle* repository [31], which is an anonymized dataset of information about women who have been diagnosed with *polycystic ovary syndrome*. The dataset consists of *541* records with *43* attributes, each corresponding to one particular patient, which includes a range of physical, hormonal, and metabolic measurements of women with *PCOS* and healthy women. Additionally, the class attribute represents the *PCOS (Y/N)* value in the dataset used to indicate whether or not an individual has *PCOS*. The primary objective of this work is to examine the prevalence of *PCOS* to distinguish between individuals who have the condition (with "*PCOS (Y/N)*" value *1*) and those who do not (with "*PCOS (Y/N)*" value *0)*. Table 1 illustrates the features of the dataset along with the data type and number of null values of each attribute.

3.2 Preprocessing

An essential step in the implementation process is data preparation [36]. The data must be modified before being included in multiple classes of algorithms that are meant to be processed and used. It can also be utilized to obtain an intelligible output in a certain format. Some of the salient steps in data preparation are as follows:

3.2.1 Data cleaning

1. ***Eliminating Extraneous Features:*** Eliminating extraneous features in the pre-processing phase refers to removing irrelevant or redundant attributes from the data before training a model. This allows the model to prioritize important features and increase the speed of the training process, leading to better pattern identification, improved model accuracy, and reduced risk of overfitting. By removing extraneous information, such as incorporating columns like "*Sl. No*" and "*Patient File No.*" from the dataset, we have improved the generalizability of our study.
2. ***Removing Missing Values:*** In real-world scenarios, it is challenging in certain situations to acquire all of the necessary information from a specific topic due to factors such as privacy concerns, disruptions in the flow of information, the patient's unwillingness to cooperate, and so on. Since the presence of missing values produces misleading results when developing models, replacing them is a crucial step in the pre-processing of data [2]. In this study, we have replaced the missing values with zero since the missing value is numerical. This approach is acceptable as it does not significantly influence the results. Furthermore, it guarantees that the classifier can continue to learn from the data since there is a relatively small number of missing values in the dataset.
3. ***Standardization of Attribute Values:*** Data normalization is utilized to standardize and make various types of data consistent, as many medical datasets consist of discrete values. One widely used data standardization method is *z-score* normalization [22], which analyzes the attribute's mean and standard deviation to normalize the data. Mathematically, it is represented as:

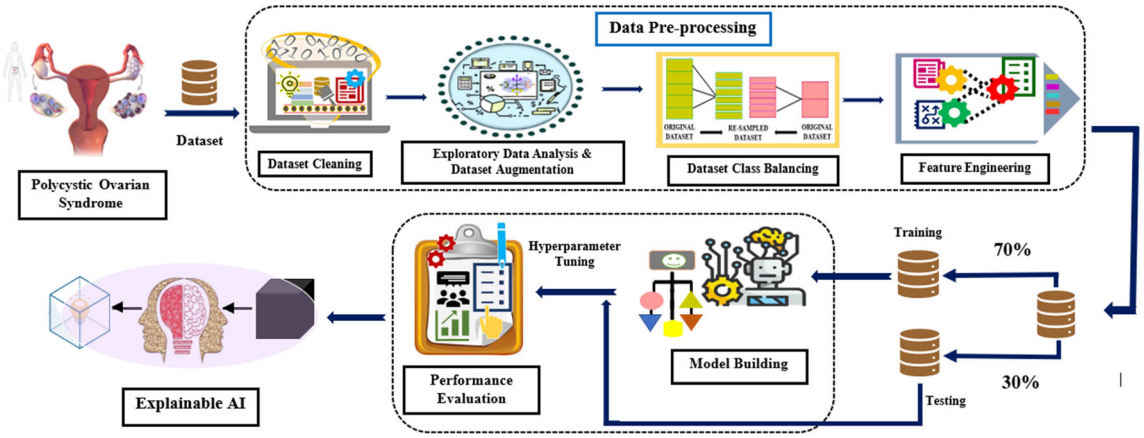


Fig. 1 Structure of a comprehensive overview of research study

Table 1 Overview of dataset: key characteristics and information

Sl. no.	Feature	Dtype	Null value	Sl. no.	Feature	Dtype	Null value
1	Sl. no.	Integer	0	23	Waist (inch)	Integer	0
2	Patient file No	Integer	0	24	Waist: hip ratio	Float	0
3	PCOS (Y/N)	Integer	0	25	Follicle No. (L)	Integer	0
4	Blood group	Integer	0	26	Follicle No. (R)	Integer	0
5	BMI	Float	0	27	Prolactin (ng/mL)	Float	0
6	Weight (Kg)	Float	0	28	Reg. Exercise (Y/N)	Integer	0
7	Cycle length (days)	Integer	0	29	RBS (mg/dl)	Float	0
8	Age (yrs)	Integer	0	30	Progesterone (ng/mL)	Float	0
9	Pulse rate (bpm)	Integer	0	31	Skin darkening (Y/N)	Integer	0
10	Height (Cm)	Float	0	32	TSH (mIU/L)	Float	0
11	Hemoglobin (Hb)	Integer	0	33	AMH (ng/mL)	Float	0
12	No. of abortions	Float	0	34	Vit D3 (ng/mL)	Float	0
13	Respiration rate	Integer	0	35	Pimples (Y/N)	Integer	0
14	I beta-HCG (mIU/mL)	Float	0	36	Avg. F size (R) (mm)	Float	0
15	II beta-HCG (mIU/mL)	Float	0	37	Hair loss (Y/N)	Integer	0
16	Pregnant (Y/N)	Integer	0	38	Avg. F size (L) (mm)	Float	0
17	Marriage Status (Yrs)	Integer	1	39	Weight gain (Y/N)	Integer	0
18	Cycle (R/I)	Integer	0	40	BP Systolic (mmHg)	Integer	0
19	LH (mIU/mL)	Float	0	41	BP Diastolic (mmHg)	Integer	0
20	FSH (mIU/mL)	Float	0	42	Hair growth (Y/N)	Integer	0
21	FSH/LH	Float	0	43	Fast food (Y/N)	Integer	1
22	Hip (inch)	Integer	0	44	Endometrium (mm)	Float	0

$$z^{score} = \frac{v - \bar{x}}{\sigma} \tag{1}$$

where v represents an individual value, $\bar{x}$ denotes the mean, and σ denotes the standard deviation of the dataset.

3.2.2 Exploratory data analysis (EDA)

Exploratory Data Analysis (EDA) [34, 38] is a technique to analyze datasets by summarizing their main characteristics, often with visual representations. *EDA* helps to identify the relationships between variables, test underlying assumptions, detect outliers and anomalies, and select appropriate models [20]. Moreover, it is often used for visualizing and

discovering insights, confirming a hypothesis, testing assumptions, and diagnosing data-related issues [26] using Python Libraries such as *NumPy, MATLAB, and Seaborn*. The *PCOS* dataset, encompassing the medical histories and clinical characteristics of individuals, has been analyzed using Histogram analysis [38] to explore the distribution of attribute values under consideration. Figure 2 displays the distribution of data by showing the number of observations that fall within each of the categories. According to Fig. 2a, both classes are more prevalent between the ages of 25 and 30. The histogram in Fig. 2b plots the Cycle (R/I) attribute and displays the maximum value of both classes when there are no irregularities in the menstruation cycle. In Fig. 2c, the histogram for Follicle No. (R) is shown. The highest count for *PCOS (yes)* is approximately 40 before the value of Follicle No. (R) 12, and at the value of 2, *PCOS* class 'No' has the highest count.

The *3D scatter plot* [9, 40] has been represented to show the relationship among three variables and also depicted how the changes in one variable affect the other two. Figure 3ay shows that *PCOS* occurs when there are more than 2 abortions and the age (in years) is under 30; while, *PCOS* does not occur when the number of abortions is less than 2 and the age (in years) is more than 30. According to Fig. 3b, the probability of *PCOS* is found to be heightened when the cycle duration (days) is less than 3 and the waist weight (kg) is greater than 100.

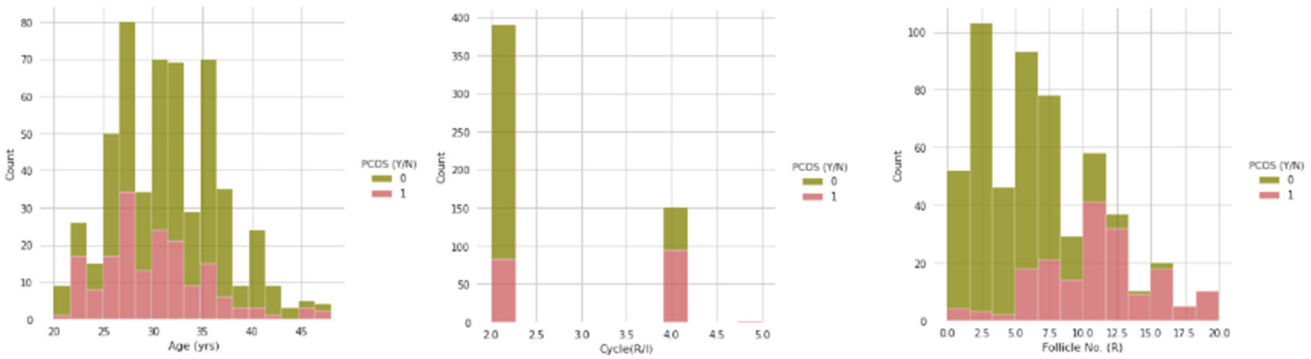
A *pair plot* [30] is a type of graphical representation that displays the relationship between several variables as a matrix of scatter plots. It is also known as a scatter plot matrix because it plots every possible pair of variables in the data set. The plots in the matrix represent the relationships between each pair of variables; while, the diagonal of the matrix is utilized to display the univariate distribution of each variable. The resulting pair plot presents a quick visual overview of the relationships between all the variables in the dataset.

In Fig. 4, the pair plot investigates each variable against the other three variables, which helps to perceive the relationships between each variable. The data can be compared with the *PCOS (Y/N)* variable to examine if certain features are more correlated with the presence or absence of *PCOS*.

3.3 Augmentation and class balancing

Overfitting is a stumbling block while predicting valid outcomes; in such a situation, a model learns the data too proficiently to correctly envisage new data. To overcome this problem, dataset augmentation has been employed in this study [27]. Augmentation is a data pre-processing technique utilized to increase the number of k-rows [42] in the dataset by creating modified versions of existing data samples, mainly when the dataset is relatively small, to increase the effectiveness of the *ML* algorithm and reduce overfitting.

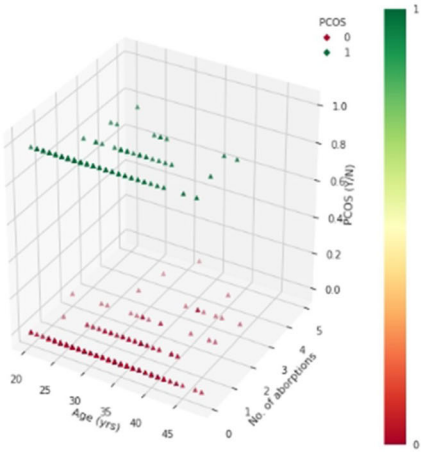
Following the data augmentation step, data balancing becomes a crucial pre-processing stage to address class imbalance issues in datasets before applying machine learning algorithms. Class imbalance is a state where one class of data is significantly more frequent than the other. This condition is effectively addressed through the hybridization of the *SMOTE* [17] and Tomek links [13], which have been executed in this study. This technique integrates the comprehension of the *SMOTE* and *Tomek Link's* [50] removal strategy to simultaneously oversample and undersample data classes. *SMOTE* works by creating new synthetic data points by taking the feature vectors of two instances from the minority class and interpolating them to increase the number of minority class samples. Conversely, *Tomek links* is a data cleaning technique that looks for pairs of instances from distinct classes that are incredibly close to each other to eliminate the instance from the majority class. Integrating both strategies together assists in mitigating imbalance in favor of the majority



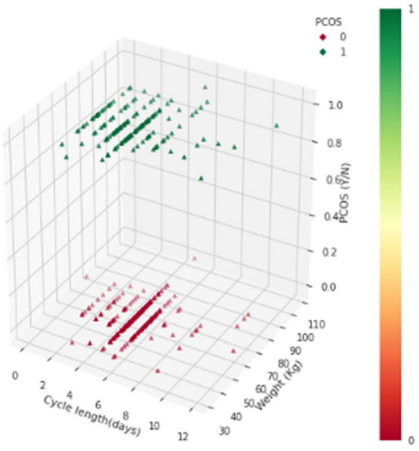
(a) Age (yrs) **(b)** Cycle (R/I) **(c)** Follicle No. (R)

Fig. 2 Assessment of the attributes in the Histogram for the PCOS (Y/N) class

Fig. 3 Examining the scattering analysis of features in 3D



(a) Age, Abortions, and PCOS (Y/N) Features for Data Visualization



(b) Cycle length (days), Weight(kg), and PCOS (Y/N) Features for Data Visualization

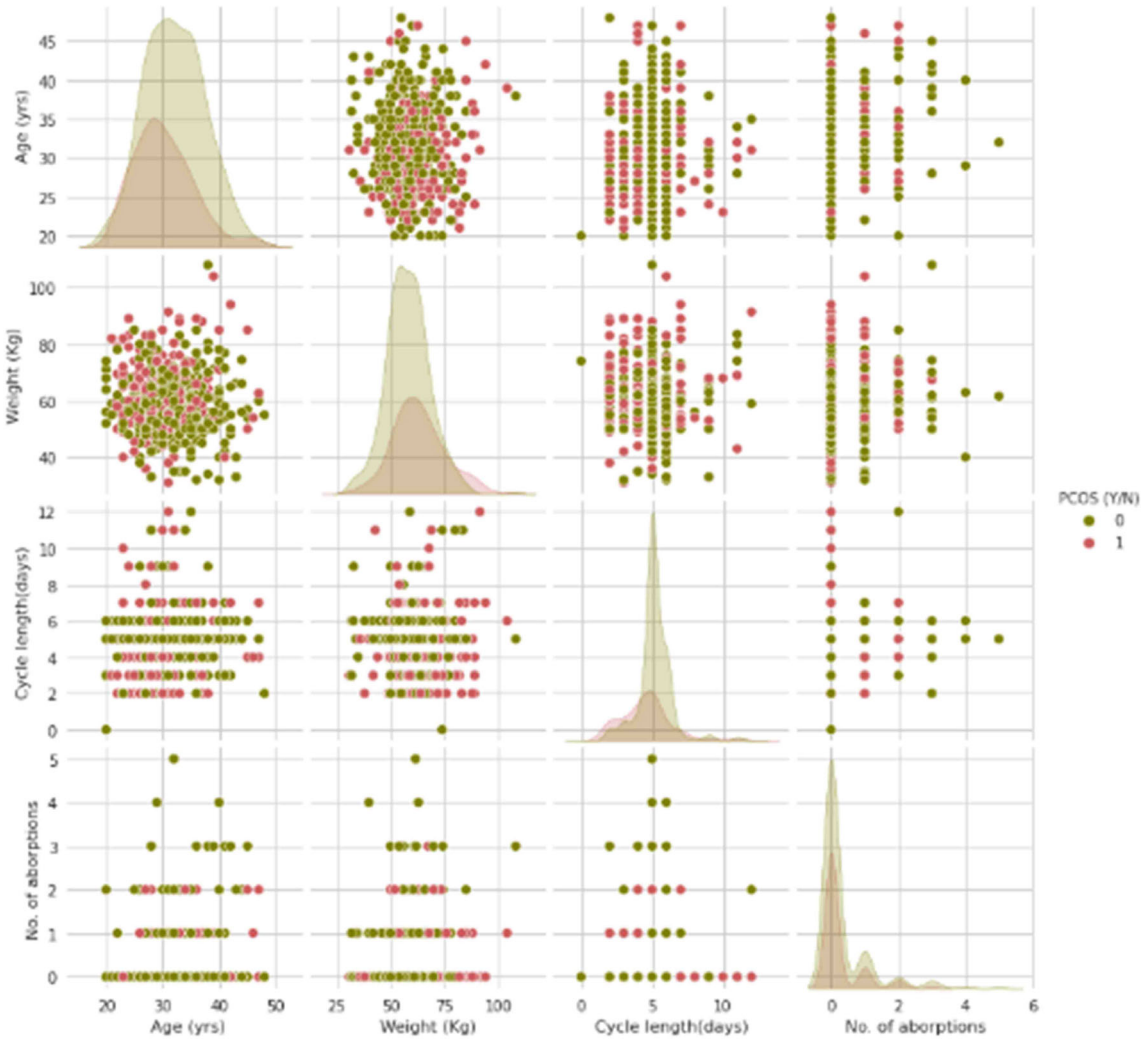


Fig. 4 Pair plot showing the relationships between age (yrs), weight (kg), cycle length (days), and no. of abortion

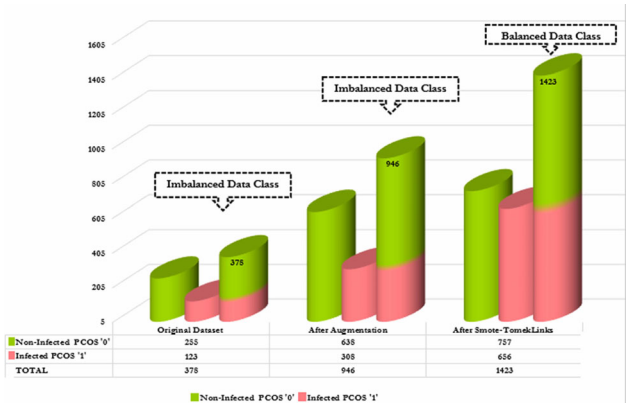


Fig. 5 Effect of SMOTE-Tomek links on data class balancing: original dataset, augmented dataset, and sampling after augmentation

class and increases the excellence of the model. Figure 5 demonstrates that the count of non-PCOS patients significantly exceeds the count of those diagnosed with the disease before using the *SMTL*. The algorithm to deal with balancing values for data is shown in Algorithm 1.

Algorithm 1 SMOTE-TomekLinks

1: $D \rightarrow$ *Original Dataset*, $A \rightarrow$ *Majority class*, $B \rightarrow$ *Minority class*
2: *Apply SMOTE to B to generate synthetic samples*
3: $synthetic^{B}_{\text{data}} \leftarrow \text{SMOTE}(data^{B})$
4: $D' \leftarrow D \cup synthetic^{B}_{\text{data}}$
5: *Initialize TOMEKLinks on* D': Identify Tomek links between A and B
6: *Remove instances in Tomek links to obtain a balanced dataset:*
7: $D^{\text{Balanced}} \leftarrow \text{RemoveTomekLinks}(D', A, B)$

3.4 Feature selection

Feature Selection (FS) [8] involves retaining the most relevant features of the dataset under investigation without losing the originality of the dataset. This is the most potential step in the data pre-processing stage, as selecting the right features can help build a better and more accurate model. The fundamental amil of feature selection is to reduce the complexity of an algorithm, which can improve its accuracy and help to mitigate the risk of overfitting. *Grey Wolf Optimization* [35] approach is a meta-heuristic technique implemented as a feature selection approach in this work that uses the concept of leadership hierarchy and the hunting behavior of grey wolves.

The social hierarchy of Grey wolves [3] is divided into the following four levels for exploring the search space and finding the best solutions:

1. ***Alpha*** (α): Alpha is the exploitation phase of the *GWO* FS process that involves randomly choosing a set of features from the dataset and evaluating their performance.
2. ***Beta*** (β): Beta is also the exploitation phase of the *GWO* FS process that involves choosing the best-performing subset of features from the dataset and then improving their performance by applying a mutation operation.
3. ***Delta*** (δ): Delta is the local search phase of the *GWO* FS process that involves choosing a subset of features from the dataset, evaluating their performance, and then performing a local search operation to enhance the effectiveness of the best-performing subset.
4. ***Omega*** (ω): Omega is the integration phase of the *GWO* FS process that combines the best-performing feature subsets from the exploitation and exploration phases to generate a final set of features. The best features are then selected from this final set.

The three main phases of gray wolves' hierarchical structure and social behavior [29] **are as follows:**

1. ***Tracking and Approaching:*** Wolves identify their prey using their keen senses of smell and sight. They will begin tracking it from a safe distance.
2. ***Encircling:*** Once the prey is surrounded, the pack leader signals the attack.
3. ***Attacking:*** Wolves launch into the attack, using their sharp teeth and claws to bring down the prey.

Mathematical Model for Gray Wolf Optimization

The fittest solution in the *GWO* mathematical model is known as *alpha* (α). *Beta* (β) and *delta* (δ) are ranked as the second and third most favorable alternatives, respectively. The remaining possible solutions are esteemed to be *omega* (ω). To hunt a prey, the pack must first encircle it. The following Eqs. (2)–(5) are utilized to model encircling behavior mathematically:

$$\vec{A}(t+1) = \overrightarrow{A_P}(t) + \vec{X} \cdot \vec{C} \tag{2}$$

$$\vec{C} = \left| \vec{Y} \cdot \overrightarrow{A_P}(t) - \vec{A}(t) \right| \tag{3}$$

where t represents the total number of iterations, C is a vector for convergence of the wolf toward the prey, A is the position of the grey wolf, and A_p is the position of the prey. X and Y are the coefficient vectors. In Eqs. (4) and (5), X and Y vectors are calculated.

$$\vec{X} = 2v\vec{r_1} - v \tag{4}$$

$$\vec{Y} = 2\vec{r_2} \tag{5}$$

where $\vec{r_1}$ and $\vec{r_2}$ represent random vectors within the range [0, 1] and $\vec{v}$ is a vector set decreasing over iterations

linearly from 2 to 0. Equation (6) is used to update the wolves' placements:

$$\overrightarrow{A}(t+1) = \frac{\overrightarrow{A_1} + \overrightarrow{A_2} + \overrightarrow{A_3}}{3} \tag{6}$$

where A_1, A_2, and A_3 are defined in Eqs. (7)–(9):

$$\overrightarrow{A_1} = \overrightarrow{A_\alpha} - X_1 \cdot (\overrightarrow{C_\alpha}) \tag{7}$$

$$\overrightarrow{A_2} = \overrightarrow{A_\beta} - X_2 \cdot (\overrightarrow{C_\beta}) \tag{8}$$

$$\overrightarrow{A_3} = \overrightarrow{A_\delta} - X_3 \cdot (\overrightarrow{C_\delta}) \tag{9}$$

where $\overrightarrow{A_\alpha}$, $\overrightarrow{A_\beta}$, and $\overrightarrow{A_\delta}$ are the wolf's highest-ranked solutions at iteration t. X_1, X_2, and X_3 are the coefficients that control the step size, and $\overrightarrow{C_\alpha}$, $\overrightarrow{C_\beta}$, and $\overrightarrow{C_\delta}$ are specified by Eqs. (10)–(12):

$$\overrightarrow{C_\alpha} = \left|\overrightarrow{Y_1} \cdot \overrightarrow{A_\alpha} - \overrightarrow{A}\right| \tag{10}$$

$$\overrightarrow{C_\beta} = \left|\overrightarrow{Y_2} \cdot \overrightarrow{A_\beta} - \overrightarrow{A}\right| \tag{11}$$

$$\overrightarrow{C_\delta} = \left|\overrightarrow{Y_3} \cdot \overrightarrow{A_\delta} - \overrightarrow{A}\right| \tag{12}$$

Ultimately, the adjustment of a parameter influences the balance between exploration and exploitation. In each iteration, the parameter undergoes a linear update, ranging from 2 to 0, as illustrated in the Eq. (13) provided below:

$$v = 2 - t \cdot \frac{2}{\mathrm{Max}^m\mathrm{Iter}^n} \tag{13}$$

where t represents the number of iterations and $\mathrm{Max}^m\mathrm{Iter}^n$ is the highest number of iterations permitted for optimization. The workflow of GWO is depicted in Fig. 6.

Figure 6 illustrates that among the pack of wolves, the alpha wolves, which possess the most desirable traits, hold a position of dominance over the beta wolves, who have less desirable traits. The beta wolves, while less important, still contribute to the decision-making process and are at risk of being replaced by the alpha wolves [29].

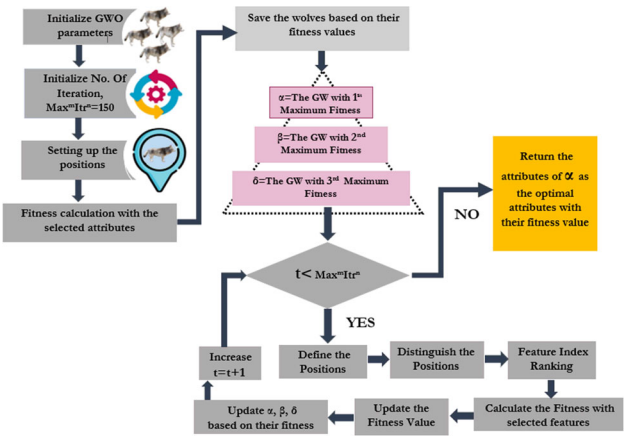


Fig. 6 A visual representation of the complete workflow of GWO for feature selection

4 Machine learning classifiers

In this Section, the operational principles and mathematical equations of seven predictive machine learning models and a recommended model for *PCOS* diagnosis are being examined.

4.1 Traditional classifiers

4.1.1 Logistic regression

Logistic regression [43, 49] is a *SML* algorithm that predicts the likelihood of an event belonging to one of the two possible classes, predicting the risk of *PCOS* disease. It accomplishes this by employing a sigmoid function or logistic function, which maps the input data to a probability value ranging from 0 to 1. The equation that used is for the logistic function is:

$$\hat{p} = \frac{e^{B_0+B_1X_1}}{1+e^{B_0+B_1X_1}} \tag{14}$$

In this equation, $\hat{p}$ signifies the estimated probability that the output variable assumes the value 1 (true) derived from the X_1. The terms B_0 and B_1 represent the intercept and coefficient associated with the input variable X_1.

4.1.2 Naive Bayes

Naïve Bayes [2, 51] is a conditional probabilistic classifier that computes a series of probability estimates determined by combining the frequency and value of a dataset. It utilizes Bayes's theorem and consider that all attributes are independent when the value of the class variable is known. This algorithm can be utilized to identify the class of a dataset and is easy to use since it does not involve much numerical optimization or matrix multiplication. Mathematically, it is calculated as:

$$P(B/D) = \frac{P(D/B) \cdot P(B)}{P(D)} \tag{15}$$

where $P(B)$ represents the independent likelihood of event B, $P(D)$ is the independent likelihood of event D, and $P(D|B)$ is the likelihood of event D given that B is true.

4.1.3 Random forest

Random Forest Classifier [2, 33, 47] is a supervised ensemble learning classifier that enhances model strength by merging multiple decision trees. This classifier operates by randomly choosing a portion of the training set and constructing a decision tree based on that subset. The final prediction is determined by combining the outcomes, achieved either through a majority vote or by averaging the predictions from each individual decision tree.

4.1.4 Adaptive boosting

The fundamental concept of ADB [41] is that each successive weak classifier places greater emphasis on the samples that were incorrectly classified by the preceding classifiers. By combining these weak classifiers, ADB can create a robust classifier that exhibits high performance on the training set and extends its effectiveness to unseen data.

4.1.5 Gradient boosting

Gradient Boosting [2, 33] is referred to as a stepwise additive model, which allows each new weak learner to be added individually while keeping all existing weak learners in the model unchanged. This approach works by constructing each weak model to predict the residuals, allowing the ensemble to gradually reduce the errors and enhance overall prediction accuracy.

4.1.6 Extreme gradient boosting

XGB [2, 10] is an optimized form of gradient boosting technique, which combines multiple weak models in an optimized and parallelized manner. It uses regularization, weighted quantile sketch, and handling of missing values to improve training speed and accuracy, resulting in a powerful ensemble model for regression, classification, and ranking tasks.

4.1.7 Categorical boosting

Categorical Boosting [24, 33] is a method that converts categorical features to numerical values and trains many decision trees in a boosting process to reduce the loss function. It uses regularization techniques and a technique called "ordered boosting" to handle categorical features and avoid overfitting. The ultimate model is created by combining the predictions generated by each individual tree.

Mathematically, the predicted probability $p_{\sigma_{t,n}}$ for category $\sigma_{t,n}$ at iteration t for data point n is expressed as follows:

$$p_{\sigma_{t,n}} = \frac{\sum_{y=1}^{t-1}[p_{\sigma_{y,n}} = p_{\sigma_{t,n}}] \cdot Q_{\sigma_y} + \beta \cdot T}{\sum_{y=1}^{t-1}[p_{\sigma_{y,n}} = p_{\sigma_{t,n}}] + \beta} \tag{16}$$

where $\sigma_{t,n}$ represents the permutation used at iteration y for data point n, Q_{σ_y} is the quality score associated with permutation σ_y at iteration y, β is the importance or weight of prior information, and T represents the prior information value.

4.2 Proposed classifier

The *Proposed Model* combines the strengths of *Boosting and Bagging* approaches to improve the performance of the ML classifier. Boosting is an EL model that integrates weak learners to construct a strong learner. Whereas, Bagging is another ensemble approach that integrates several models trained on different subsets of the data to reduce overfitting. The proposed classifier is mainly partitioned into two phases: In the first phase, this approach employs various boosting classifiers such as *GB, ADB, XGB, CATB* along with *LR* and a bagging classifier like

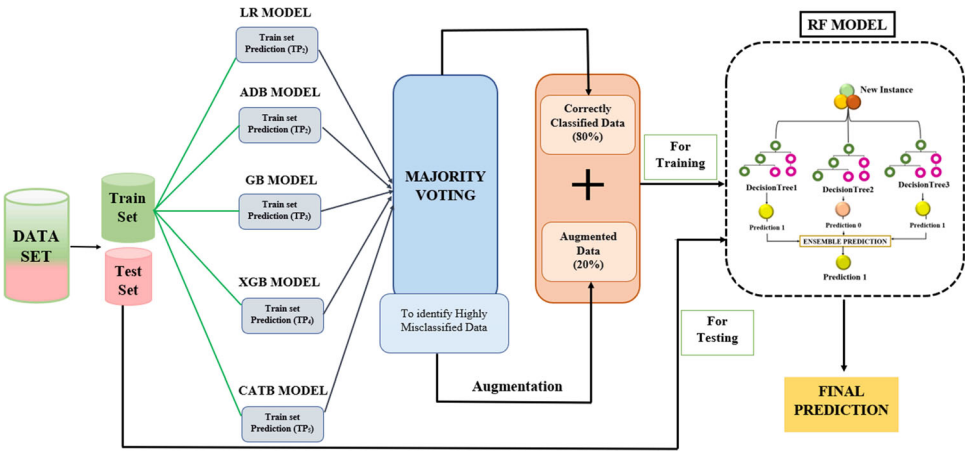


Fig. 7 Detecting PCOS with machine learning: a proposed model

Random Forest (RF) to build a blended ensemble model. The original dataset is splitted into a training dataset and a testing dataset, with 70% of the dataset being used for training and the remaining 30% for testing. The PODBoost model trains multiple boosting classifiers with different parameters on the training data to combine weak models into stronger one. It computes predicted values for the training tuples and identifies maximally misclassified tuples. Training accuracy is measured in this phase. In the second phase, 80% of the correctly predicted observations and 20% of the maximally misclassified data are considered to train a *Random Forest* classifier. Here the proposed method used a weighted approach to select the majority of misclassified observations. These combined predictions are then used to train a *RF* classifier, which generates final predictions for the testing set. The pseudocode for the PODBoost model is explained in Algorithm 2 (Fig. 7).

Algorithm 2 PODBoost Model

1: **Algorithm:** $D \rightarrow$ Input Dataset
2: Split D into $T \rightarrow$ Training Set and $V \rightarrow$ Testing Set
3: Model training and hyperparameter tuning on T using LR, ADB, GB, XGB, and CATB
4: For each classifier, compute classification performance metrics for T
5: Execute voting scheme to identify maximally misclassified tuples Y
6: Create modified training dataset T_{new} by blending 80% correctly predicted tuples and 20% highly misclassified tuples
7: Train model M using the Random Forest algorithm on T_{new}
8: Use the trained model M for the final prediction of test data V

4.3 Hyperparameter tuning

Finding the best values for the adjustable parameters of an ML model to achieve optimal performance is known as hyperparameter tuning [15]. To tune the hyperparameters, multiple training sessions have to be run with different hyperparameter values, and the best model performance must be chosen from the results of those training sessions. Among the various techniques for hyperparameter tuning,*Grid search* stands out as one of the simplest and most widely adopted methods, which has been used in this work. The idea behind grid search [23] is to define a grid of hyperparameter values and then search exhaustively through this grid identify the best set of hyperparameter values. Table 2 represents the analyses of the hyperparameter tuning of our research models.

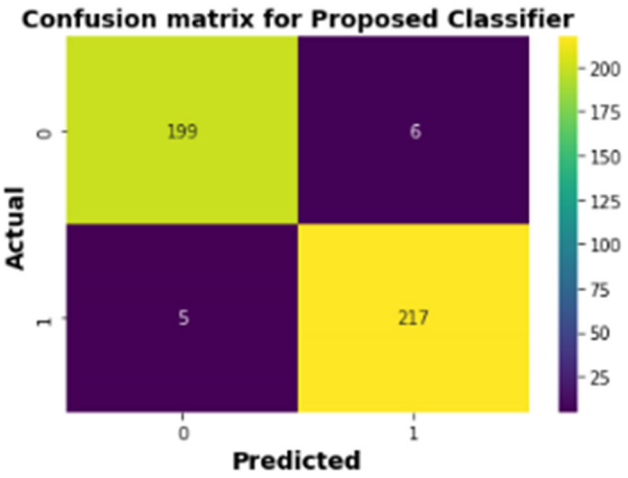


Fig. 8 Representation of confusion matrix

5 Performance evaluation measures and experimental results

This section evaluates and compares the effectiveness of the suggested *PCOS* detection framework to the proposed and other popular traditional ML models with an optimized feature set using several performance measures.

Table 2 Grid search hyperparameter analysis of implemented machine learning classifiers

Algorithm	Hyperparameters
LR	`solver='liblinear', penalty='l1', C=10`
NB	`var_smoothing=1e-8`
RF	`bootstrap=True, n_estimators=200, min_samples_leaf=4, min_samples_split=10, criterion='entropy', max_depth=70`
ADB	`learning_rate=0.1, n_estimators=100`
GB	`learning_rate=0.1, max_depth=8, min_samples_split=500, min_samples_leaf=50, subsample=0.8`
XGB	`learning_rate=0.1, subsample=0.8, max_depth=3, colsample_bytree=0.8, n_estimators=100`
CATB	`iterations=200, learning_rate=0.01, depth=10, eval_metric='Accuracy', random_seed=42`

Table 3 Performance comparison of ML algorithms without suggested method

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)	Error-rate (%)
LR	80.36	71.42	66.03	68.62	76.66	19.64
NB	85.27	83.72	67.92	74.99	84.69	14.73
RF	89.57	91.11	75.93	82.83	85.42	10.43
ADB	88.34	85.41	77.35	81.18	85.49	11.66
GB	89.70	88.00	81.48	84.61	87.99	10.3
XGB	90.18	91.30	77.77	84.00	87.05	9.82
CATB	91.41	93.47	79.62	85.99	88.43	8.59

5.1 Performance evaluation metrics

To assess the performance of *SML* in predicting *PCOS*, numerous metrics such as *Accuracy, Error-rate, Precision, AUC, Recall,* and *F1-Score* are employed. These performance indicators are beneficial in assessing the efficiency of machine learning techniques. Precision, A positive classification is applied when an individual is identified as having *PCOS* problem; while, a negative classification is provided when the individual does not suffer from *PCOS*. The confusion matrix (CM) [28, 44] of the *proposed algorithm* is represented in Fig. 8 and is used as a tool to analyze the performance of the *proposed classifier*, which comprises four major components:

- **True Positive (TPos):** When a test outcome accurately indicates the presence of PCOS in an individual.
- **False Positive (FPos):** When a test outcome inaccurately indicates the presence of PCOS in an individual.
- **True Negative (TNeg):** When a test outcome accurately indicates the absence of PCOS in an individual.
- **False Negative (FNeg):** When a test outcome inaccurately indicates the absence of PCOS in an individual.

Without utilizing the suggested method, Table 3 represents a comparison of the performance of used ML classifiers using various performance metrics. The following are the fundamental metrics for evaluation:

1. **Accuracy:** Accuracy [28, 44] is a measure of how accurately a model is able to predict a target class. The accuracy of a model is also related to its error rate. Less error occurs when accuracy is higher. Mathematically, it is represented as:

$$\text{Accuracy (Acc)} = \frac{TPos + TNeg}{TPos + FPos + TNeg + FNeg} \tag{17}$$

$$\text{Error (Err)} = \frac{FPos + FNeg}{TPos + FPos + TNeg + FNeg} \tag{18}$$

2. **Precision:** Precision [28, 44] is the fraction of true positives out of all the positive predictions made by the classifier. Mathematically, it is calculated by:

$$\text{Precision (Pre)} = \frac{TPos}{TPos + FPos} \tag{19}$$

3. **Recall:** Recall [28, 44] is a measure of the ability of a model to detect positive examples in a dataset accurately. It is usually measured by the ratio of positive examples that were accurately identified. Mathematically:

$$\text{Recall (Re)} = \frac{TPos + TNeg}{TPos + FNeg} \tag{20}$$

4. **F1-Score:** F1-Score [28, 44] is a more comprehensive measure than accuracy, as it takes into account the precision and recall of a model. Mathematically, it is computed as:

$$\text{F1-Score} = \frac{2 \times \text{Pre} \times \text{Re}}{\text{Pre} + \text{Re}} \tag{21}$$

5. **ROC-AUC:** AUC-ROC curve [28, 44] is a measurement tool for classification problems at various threshold levels. AUC stands for the amount of separability, and ROC is a graph that shows the capability of the model to distinguish between different categories.

5.2 Experimental result analysis

The 8th generation Intel Core i5, RAM-8-GB unit, and Python programming tools have been used to implement the suggested framework in steps. We have explored using machine learning techniques such as *LR, NB, RF, ADB, GB, XGB, CATB*, and the *proposed classifier* for predicting *Polycystic Ovary Syndrome (PCOS)* using a dataset collected from the Kaggle repository. The dataset has 541 data points and 43 attributes where missing values have been replaced with zeroes. In the pre-processing stage, the dataset features have been reduced to 41 columns by omitting the two additional features, "*Sl. No.*" and "*Patient File No.*" Due to the limited size of the dataset, overfitting may occur during the training process. We have applied data augmentation techniques to mitigate this issue, which involves generating additional training data to the

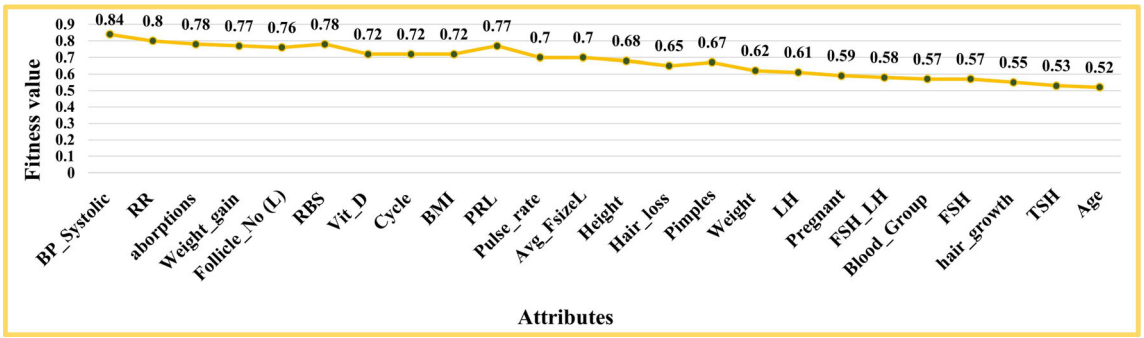


Fig. 9 Feature fitness value: selection for research study

existing data samples. This increases the number of rows or instances in the dataset, thereby allowing the classifier to learn more robust features and reducing the risk of overfitting.

To ensure reliable predictions while predicting PCOS from the dataset, a powerful pre-processing sampling approach that combines SMOTE and Tomek Links has been implemented. This approach ensures that the data is balanced, preventing the model from over-predicting positive cases or under-predicting negative cases. As a result, we ensure that our model is trained on a balanced dataset, which improves its ability to accurately predict both positive and negative cases of PCOS disease.

Moreover, we have employed feature selection using the Grey Wolf Optimization approach, followed by dataset splitting using a 70% and 30% ratio for training and testing, respectively, to increase the computational efficiency of the classification process by allowing the algorithms to run on the training set rather than the entire dataset. After that, the test set is fed into the model to test how effective the algorithm is at learning. A *GWO* algorithm has been employed to decrease the number of characteristics from *41* to *25* to minimize the computational burden. The fitness value of selected features is represented in Fig. 9, where the green color indicates the feature that has been used in this research study, and the brown color indicates the feature that has not been used. Based on their fitness value, features are chosen depending on how accurately they are relevant to the study and how well they can add to the findings.

Accuracy, precision, recall, AUC score, F1-score, and *error rate* are the six most common performance metrics used with hyperparameter tuning to evaluate the effectiveness of each classifier. This allows for a comprehensive analysis of each model's performance and helps to determine the best one for detecting *PCOS*.

Table 4 compares the performance of the proposed approach, which is a modified version of a tree-based ensemble model, with that of other tree-based models (i.e., *GB, ADB, RF, XGB, and CATB*) as well as with two different machine learning conventional models (i.e., LR and NB), with various numbers of features. The findings of the study suggest that the proposed *PODBoost* classifier outperforms others with selected features, highlighting the importance of feature selection and choosing an efficient classifier for accurate *PCOS* prediction.

As demonstrated in Fig. 10, the suggested method outperforms all conventional ML algorithms in terms of *precision, accuracy, recall, F1-Score and ROC-AUC*, validating its efficacy and relevance, while Fig. 11 illustrates the accuracy and error rate of various classifiers, with each providing different levels of accuracy depending on the data. Figure 12 compares the *ROC-AUC* score curves of machine learning models with the proposed technique for evaluation. Our results demonstrate that the *GWO-based* feature selection approach significantly improves the performance of ML classifiers compared to using all available features. In contrast to many prior studies in the realm of *PCOS* classification, our research introduces a multifaceted approach that significantly advances existing methodologies, as depicted in Table 5. While conventional studies often employ simplistic feature selection methods [5–7, 48, 53] or lack transparency in model predictions, our methodology utilizes *Grey Wolf Optimization* for feature selection, facilitating the identification of the most discriminative features for *PCOS* classification. This optimization algorithm surpasses traditional approaches by dynamically adapting to the characteristics of the dataset. Moreover, we introduce a novel hybrid classifier, *PODBoost*, which effectively integrates multiple classification algorithms to achieve superior performance in distinguishing between *PCOS* and *non-PCOS* patients. Additionally, our study incorporates *LIME*, an *XAI* method, to interpret the predictions made by our hybrid classifier. This aspect of our approach is particularly prominent as most prior research in this field often overlooks the incorporation of explainable AI techniques [18, 21, 37, 46]. The utilization of explainable AI facilitates the interpretability of the predictions made by the proposed model. By leveraging *LIME*, we provide transparent insights into the decision-

Table 4 Feature selection impact on machine learning model performance: a comparative evaluation with unseen test data

Models	Features	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LR	30	91.46	97.21	84.87	90.63
	25	**95.31**	**95.91**	**95.04**	**95.48**
	20	93.6	94.11	98.75	93.43
	15	90	87.01	94.37	90.54
	10	86.49	84.26	90.82	87.42
NB	30	93.57	91.51	96.24	93.82
	25	**95.6**	**96.34**	**95.01**	**95.67**
	20	94.07	94.6	93.24	93.92
	15	92.41	91.38	93.17	92.27
	10	91.46	92.92	90.36	91.62
RF	30	94.29	96.55	92.02	94.23
	25	**95.78**	**95.95**	**95.95**	**95.95**
	20	94.79	95.12	94.2	94.67
	15	92.65	93.94	90.73	92.31
	10	92.65	93.08	92.66	92.87
ADB	30	94.52	94.81	94.36	94.58
	25	**96.25**	**95.98**	**96.84**	**96.48**
	20	94.31	96.92	91.3	94.03
	15	92.89	91.86	93.65	92.741
	10	91.7	95.97	87.61	91.6
GB	30	95.47	94.9	96.24	95.6
	25	**96.7**	**96.42**	**97.3**	**96.86**
	20	95.49	97	93.71	95.33
	15	92.9	94.42	90.73	92.53
	10	92.9	94.76	91.28	93
XGB	30	95.47	95.32	95.77	95.55
	25	**96.96**	**97.7**	**96.4**	**97.05**
	20	95.73	96.56	94.69	95.61
	15	93.6	92.38	94.63	93.49
	10	93.13	93.15	93.57	93.36
CATB	30	95.71	94.93	96.71	85.81
	25	**97.2**	**96.89**	**97.47**	**97.32**
	20	95.97	96.57	95.17	95.86
	15	93.84	93.66	93.66	93.66
	10	93.6	94	93.57	93.79
PODBoost	30	96.2	95.4	97.18	96.28
	25	**97.42**	**97.32**	**97.49**	**97.39**
	20	96.45	97.52	95.16	96.93
	15	94.85	97.28	93.04	95.11
	10	93.84	94.03	94.03	94.03

making process of our model, enhancing the interpretability and reliability of our classification results. In particular, our approach achieved a classification accuracy of 97.42%, recall of 97.49%, and precision of 97.32% using a proposed *(PODBoost)* Classifier.

5.3 Model interpretability through LIME

The use of ML classifiers for the prediction and classification of observations of the PCOS dataset has been successful. However, the lack of transparency in the rationale behind the decision-making of these models has caused difficulty for people to trust their conclusions. To control this situation, interpretability and Explainable AI [14] were

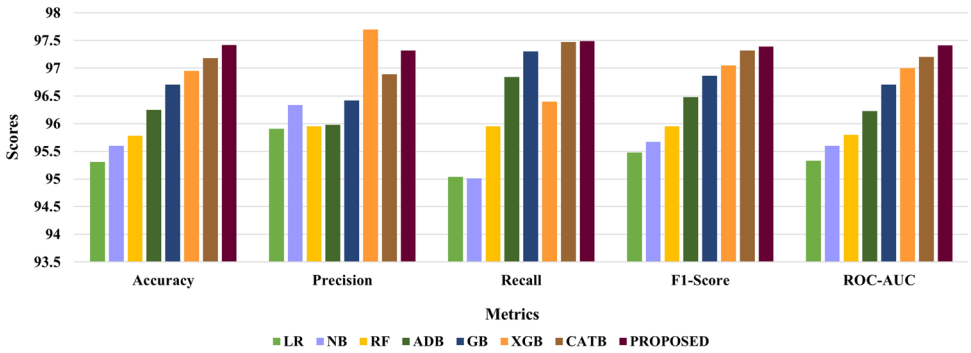


Fig. 10 Performance evaluation of machine learning models: a comparative approach using proposed technique

Fig. 11 Classifier accuracy and error rates: a comparative analysis

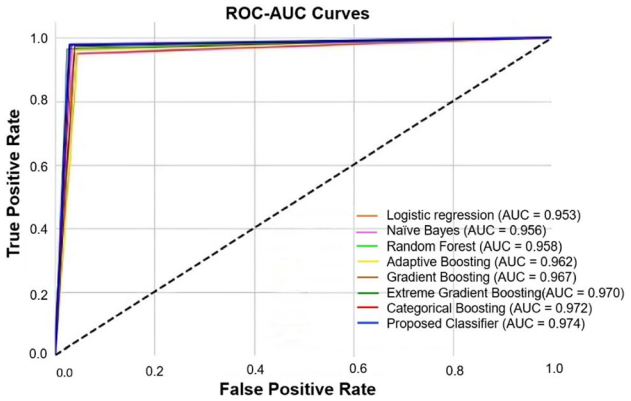


Fig. 12 ROC-AUC score comparison of machine learning models with proposed model

developed to elucidate the underlying reasons behind a model's prediction. Interpretability is the degree to which a person is able to recognize the reasoning behind the output of a "black box" model [45].

In this work, we have implemented *Local Interpretable Model-Agnostic Explanations (LIME)* [19] to discover an easily understandable model over an interpretable representation that is regionally faithful to the model or to describe the predictions that the model produced. *LIME* treats the model as a black box, which allows it to explain any model and will also work with future classifiers. This framework aims to take a predictive model and create an interpretable version. To do this, we first constructed the model for prediction and then fed it into the LIME algorithm to generate an interpretable model. This interpretable model provides text-based and visual explanations, revealing the parameters' impact on the output, whether positive or negative as shown in Fig. 13.

Additionally, two essential criteria must be met: The explanation must be easy to understand, even by those without *machine learning* knowledge, and it must exhibit local fidelity, meaning it must align with how the classifier responds near the instance it is predicting. Algorithm 3 outlines the operational steps and methodology employed by *LIME*

The prediction probabilities of the PODBoost model are displayed in the left portion of Fig. 13 for the classes "PCOS(Yes)" (100%) and "PCOS(No)" (0%) for the fifteenth observation of the dataset. The middle section displays the features and their corresponding weight that affect the probability of the prediction. For example, we can see if an HB value between 33.00 and 37.00 contributes to not getting PCOS. The importance of a certain attribute for a specific location is illustrated in the right portion.

Algorithm 3 Model Interpretability through *LIME*

1: Select the instance to be explained.
2: Generate a set of perturbations around the instance by randomly sampling from a predefined distribution.
3: Use the perturbed instances to obtain predictions from the original model.
4: Fit a simpler or interpretable model to the perturbed instances and their corresponding predictions.
5: Utilise the simpler model to illustrate the prediction of the original model for the instances being explained.

6 Conclusion and future work

In conclusion, *Polycystic Ovarian Syndrome* is a multifaceted health issue that poses a significant diagnostic challenge to medical practitioners for its accurate diagnosis. However, by leveraging machine learning techniques,

Table 5 Comparative analysis of methodologies used in previous studies

Author	Year	Samples	Sampling	Feature engineering	Machine learning classifiers	Maximum accuracy (%)	XAI
Tiwari et al. [48]	2022	541	SMOTE	Pearson's Correlation Coefficient	SVM, LR, RF, ADB, DT, K-NN, GB, XGB, CATB, LDA & QDA	93.25	NA
Zigarelli et al. [53]	2022	541	NA	Principle Component Analysis	Categorical Boosting (CATB)	90.10	SHAP
Bhardwaj et al. [7]	2021	541		Pearson's Correlation Coefficient	SVM, DT, RF, XGB & MLP	93	NA
Bharati et al. [6]	2020	541	NA	Filtering based univariate	GB, RF, LR & Hybrid	91.01	NA
Subha et al. [46]	2024	541	SMOTENC	Swarm intelligence, Flashing Firefly & Particle Swarm Optimization	RF & XGB	92.64	NA
Faris et al. [18]	2023	541	NA	Genetic Algorithm	SVM, DT, NB & KNN	89.51	NA
Batra et al. [5]	2023	541	SMOTE	Correlation-based	LR, RF & SVM	92.02	NA
Nandipati et al. [37]	2020	541	SMOTE	SelectKBest, Backward elimination, Forward selection, Correlation matrix & Chi2	KNN, SVM, RF, NB & MLP & Ensemble	93.12	NA
Gupta et al. [21]	2022	541	NA	NA	LDA & QDA	97.37	NA
Proposed algorithm (PODBoost)	–	541	SMTL	Grey Wolf optimization	Proposed classifier	97.42	LIME

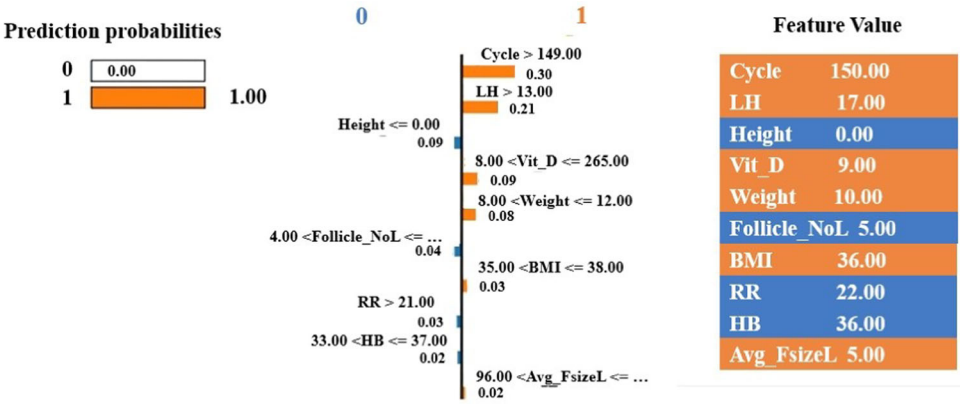


Fig. 13 An explanation based on the likelihood of individuals being PCOS (Yes) or PCOS (No) by the proposed classifier

we have demonstrated the potential to reliably predict the presence of this condition using relevant clinical features. In this work, data has been collected from *Kaggle* and augmented to increase the number of rows. We have also tried to balance the data using *Smote-TomekLinks*, and relevant features have been selected using *Grey Wolf optimization*. Several classifiers have been trained, including *Random Forest, Adaptive Boosting, Naive Bayes, logistic regression, Gradient Boosting, Extreme Gradient Boosting, categorical boosting, and one proposed hybrid classifier*. The performance of these classifiers has been evaluated utilizing numerous metrics, including *accuracy, ROC-AUC score, recall, precision, F1-score, and error rate* after hyperparameter tuning. During the study, it has been recognized that the proposed classifier works very well in identifying *PCOS*, providing a promising framework for predicting *PCOS* with an accuracy of 97.42%. Moreover, future work could include the utilization of multi-modality datasets in the diagnosis of *PCOS*, such as ultrasound scans, and using distinct or more extensive real-

life datasets for diagnosis. Additionally, we are working on different datasets to predict other similar kinds of diseases using the proposed algorithm and expect that the proposed algorithm could provide us with better prediction results.

Acknowledgements This research work is part of the project supported by DST under the PURSE 2022 scheme.

Data availability The dataset used in this study is publicly accessible on Kaggle and can be found at the following link: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos.

Declarations

Conflict of interest The author affirms the absence of any Conflict of interest and asserts that there are no financial resources involved.

References

1. Aggarwal S, Pandey K (2023) Early identification of pcos with commonly known diseases: obesity, diabetes, high blood pressure and heart disease using machine learning techniques. Expert Syst Appl 217:119532
2. Ahamed BS, Arya MS (2022) Nancy AOV (2022) Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation. Adv Hum-Comput Interact 1:9220560
3. Al-Tashi Q, Rais H, Jadid S (2019) Feature selection method based on grey wolf optimization for coronary artery disease classification. In: Recent trends in data science and soft computing: proceedings of the 3rd international conference of reliable information and communication technology (IRICT 2018), Springer, pp 257–266
4. Artini PG, Obino MER, Sergiampietri C et al (2018) Pcos and pregnancy: a review of available therapies to improve the outcome of pregnancy in women with polycystic ovary syndrome. Expert review of endocrinology & metabolism 13(2):87–98
5. Batra H, Nelson L (2023) Dcads: Data-driven computer aided diagnostic system using machine learning techniques for polycystic ovary syndrome. International Journal of Performability Engineering 19(3)
6. Bharati S, Podder P, Mondal MRH (2020) Diagnosis of polycystic ovary syndrome using machine learning algorithms. In: 2020 IEEE region 10 symposium (TENSYMP), IEEE, pp 1486–1489
7. Bhardwaj P, Tiwari P (2022) Manoeuvre of machine learning algorithms in healthcare sector with application to polycystic ovarian syndrome diagnosis. In: Proceedings of Academia-Industry Consortium for Data Science: AICDS 2020. Springer, p 71–84
8. Cai J, Luo J, Wang S et al (2018) Feature selection in machine learning: A new perspective. Neurocomputing 300:70–79
9. Casa A, Scrucca L, Menardi G (2021) Better than the best? answers via model ensemble in density-based clustering. Adv Data Anal Classif 15:599–623
10. Choi DK (2019) Data-driven materials modeling with xgboost algorithm and statistical inference analysis for prediction of fatigue strength of steels. Int J Precis Eng Manuf 20:129–138
11. Çiçek İB, Küçükakçali Z, Yağin FH (2021) Detection of risk factors of pcos patients with local interpretable model-agnostic explanations (lime) method that an explainable artificial intelligence model. J Cognit Syst 6(2):59–63
12. Danaei Mehr H, Polat H (2022) Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. Health Technol 12(1):137–150
13. Devi D, Biswas SK, Purkayastha B (2019) Learning in presence of class imbalance and class overlapping by using one-class svm and undersampling technique. Connect Sci 31(2):105–142
14. Duell J, Fan X, Burnett B, et al (2021) A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records. In: 2021 IEEE EMBS international conference on biomedical and health informatics (BHI), IEEE, pp 1–4
15. Elgeldawi E, Sayed A, Galal AR, et al (2021) Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In: Informatics, MDPI, p 79
16. Elmannai H, El-Rashidy N, Mashal I et al (2023) Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence. Diagnostics 13(8):1506
17. Elreedy D, Atiya AF (2019) A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. Inf Sci 505:32–64
18. Faris NN, Miften FS (2023) Proposed model for detection of pcos using machine learning methods and feature selection. J Educ Pure Sci-Univ Thi-Qar 13(1):85–93
19. Gabbay F, Bar-Lev S, Montano O et al (2021) A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients. Appl Sci 11(21):10417
20. Ganie SM, Malik MB (2022) An ensemble machine learning approach for predicting type-ii diabetes mellitus based on lifestyle indicators. Healthc Anal 2:100092
21. Gupta A, Soni H, Joshi R, et al (2022) Discriminant analysis in contrasting dimensions for polycystic ovary syndrome prognostication. arXiv preprint arXiv:2201.03029
22. Henderi H, Wahyuningsih T, Rahwanto E (2021) Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. Int J Inf Inf Syst 4(1):13–20
23. Hoque KE, Aljamaan H (2021) Impact of hyperparameter tuning on machine learning models in stock price forecasting. IEEE Access 9:163815–163830
24. Huang G, Wu L, Ma X et al (2019) Evaluation of catboost method for prediction of reference evapotranspiration in humid regions. J Hydrol 574:1029–1041
25. Hussain S, Mustafa MW, Jumani TA et al (2021) A novel feature engineered-catboost-based supervised machine learning framework for electricity theft detection. Energy Reports 7:4425–4436
26. Indrakumari R, Poongodi T, Jena SR (2020) Heart disease prediction using exploratory data analysis. Proc Comput Sci 173:130–139
27. Inoue H (2018) Data augmentation by pairing samples for images classification. arXiv preprint arXiv:1801.02929
28. Jiao Y, Du P (2016) Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant Biol 4:320–330
29. Kamel SR, YaghoubZadeh R, Kheirabadi M (2019) Improving the performance of support-vector machine by selecting the best features by gray wolf algorithm to increase the accuracy of diagnosis of breast cancer. J Big Data 6:1–15
30. Khare V, Kumari S (2022) Performance comparison of three classifiers for fetal health classification based on cardiotocographic data. Acadlore Trans AI Mach Learn 1(1):52–60
31. Kottarathil P (2020) Polycystic ovary syndrome (pcos) dataset. https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos
32. Mathur P, Kakwani K, Diplav, et al (2020) Deep learning based quantification of ovary and follicles using 3d transvaginal

ultrasound in assisted reproduction. In: 2020 42nd annual international conference of the ieee engineering in medicine and biology society (EMBC), pp 2109–2112
33. Mienye ID, Sun Y (2022) A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access 10:99129–99149
34. Milo T, Somech A (2020) Automating exploratory data analysis via machine learning: An overview. In: Proceedings of the 2020 ACM SIGMOD international conference on management of data, pp 2617–2622
35. Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. Adv Eng Softw 69:46–61
36. Misra P, Yadav AS (2019) Impact of preprocessing methods on healthcare predictions. In: Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE)
37. Nandipati S, Ying C, Wah KK (2020) Polycystic ovarian syndrome (pcos) classification and feature selection by machine learning techniques. Appl Math Comput Intell 9:65–74
38. Nasim S, Almutairi MS, Munir K et al (2022) A novel approach for polycystic ovary syndrome prediction using machine learning in bioinformatics. IEEE Access 10:97610–97624
39. Patel S (2018) Polycystic ovary syndrome (pcos), an inflammatory, systemic, lifestyle endocrinopathy. J Steroid Biochem Mol Biol 182:27–36
40. Pfister L, Wetzel CE, Klaus J et al (2017) Terrestrial diatoms as tracers in catchment hydrology: a review. Wiley Interdiscip Rev Water 4(6):e1241
41. Rahmani AM, Shafique M, Jantsch A et al (2018) adboost: Thermal aware performance boosting through dark silicon patterning. IEEE Trans Comput 67(8):1062–1077
42. Sagadeeva S, Boehm M (2021) Sliceline: fast, linear-algebra-based slice finding for ml model debugging. In: Proceedings of the 2021 international conference on management of data, pp 2290–2299
43. Schein AI, Ungar LH (2007) Active learning for logistic regression: an evaluation. Mach Learn 68:235–265
44. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manage 45(4):427–437
45. Srinivasu PN, Sandhya N, Jhaveri RH et al (2022) From blackbox to explainable AI in healthcare: existing tools and case studies. Mobile Inf Syst 2022:1–20
46. Subha R, Nayana B, Radhakrishnan R et al (2024) Computational intelligence for early detection of infertility in women. Eng Appl Artif Intell 127:107400
47. Talukdar S, Eibek KU, Akhter S et al (2021) Modeling fragmentation probability of land-use and land-cover using the bagging, random forest and random subspace in the teesta river basin, bangladesh. Ecol Ind 126:107612
48. Tiwari S, Kane L, Koundal D et al (2022) Sposds: a smart polycystic ovary syndrome diagnostic system using machine learning. Expert Syst Appl 203:117592
49. Tsangaratos P, Ilia I (2016) Comparison of a logistic regression and naïve bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. Catena 145:164–179
50. Wang Z, Wu C, Zheng K et al (2019) Smotetomek-based resampling for personality recognition. IEEE Access 7:129678–129689
51. Yang FJ (2018) An implementation of naive bayes classifier. In: 2018 International conference on computational science and computational intelligence (CSCI), IEEE, pp 301–306
52. Zhang XZ, Pang YL, Wang X et al (2018) Computational characterization and identification of human polycystic ovary syndrome genes. Sci Rep 8(1):12949
53. Zigarelli A, Jia Z, Lee H (2022) Machine-aided self-diagnostic prediction models for polycystic ovary syndrome: observational study. JMIR Format Res 6(3):e29967