Intelligence-Based Medicine 11 (2025) 100253

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine

Enhanced Polycystic Ovary Syndrome diagnosis model leveraging a K-means based genetic algorithm and ensemble approach

Najlaa Faris [a], Aqeel Sahi [b,c,*], Mohammed Diykh [d,e], Shahab Abdulla [e], Siuly Siuly [f]

[a] Southern University of Technology, Thi-Qar Technology College, Thi-Qar, 64001, Iraq
[b] School of Mathematics, Physics and Computing, University of Southern Queensland, QLD, 4350, Australia
[c] College of Engineering, Al-Shatrah University, Thi-Qar, 64001, Iraq
[d] Technical Engineering College, Department of Cybersecurity, Al-Ayen Iraqi University, Thi-Qar, 64001, Iraq
[e] UniSQ College, University of Southern Queensland, QLD, 4350, Australia
[f] Institute for Sustainable Industries & Livable Cities, Victoria University, VIC, 3000, Australia

ARTICLE INFO

Keywords:
Polycystic ovary syndrome
Detection
Ensemble methods
Genetic algorithm
K-means

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is a prevalent hormonal disorder affecting women in their childbearing years. Detecting PCOS early is crucial for preserving fertility in young women and preventing long-term health complications like hypertension, heart disease, and obesity. While costly clinical tests exist to detect PCOS, there is a growing demand for more accurate and affordable diagnostic methods. The primary objective of this research is to pinpoint the most effective PCOS features that can aid experts in early diagnosis. We introduce a feature extraction model, termed KM-GN, which combines the k-means algorithm with a genetic selection algorithm to identify the most informative features for PCOS detection. These selected features are fed into our designed model, Random Subspace-based Bootstrap Aggregating Ensembles (RSBE). To assess the performance of the proposed RSBE method, we compare it against several individual and ensemble classifiers. The effectiveness of our model is assessed using a freely accessible dataset comprising 43 traits from 541 women, of whom 177 have been diagnosed with PCOS. We employ various statistical metrics to evaluate the performance, including the confusion matrix, accuracy, recall, F1 score, precision, and specificity. The experimental outcomes demonstrate the viability of implementing our proposed model as a hardware tool for efficient detection of PCOS.

1. Introduction

Polycystic ovary syndrome (PCOS) is a hormone disorder that affects women. Clinical studies have shown that PCOS can be distinguished by hyperandrogenism [1–6]. Stein and Leventhal first diagnosed this disease in 1935 [2]. Women with PCOS often show some symptoms, such as menstrual and infertility issues [3]. Moreover, some women may suffer from long-term health issues such as heart disease, diabetes, mood disorders, and uterine cancer [4]. The exact causes of PCOS are still not identified. However, many clinical studies have shown that genetic factors or excess androgen and insulin resistance may play a vital role in developing PCOS.

Irregular or absent periods, body hair growth, acne, scalp hair loss, and high levels of testosterone are considered the common signs of PCOS. Those symptoms and signs may vary in severity among women, ranging from mild to severe, making the condition difficult to diagnose [7].

Experts usually use clinical data or ultrasound scans to diagnose PCOS; however, less than 50 % of women are correctly diagnosed and receive the proper treatment [8]. The literature has noticed that clinical data, such as obesity, heart disease, high blood pressure, diabetes, etc., are widely used in PCOS diagnosis compared with ultrasound images due to their availability and complexity.

Many recent studies based on machine learning approaches have shown that there is a relationship between the development of PCOS and obesity, high blood pressure, heart disease, and diabetes. For example, Aggarwal et al. (2023) designed a machine learning-based model to diagnose PCOS. A data amalgamation with feature selection methods was suggested, resulting in 8 parameters and 985 records. Rahman et al., 2024 integrated Mutual Information with decision tree, AdaBoost, random forest, logistic regression, decision tree, AdaBoost, random forest, and support vector machine. Mutual Information was applied for

* Corresponding author. School of Mathematics, Physics and Computing, University of Southern Queensland, QLD, 4350, Australia.
E-mail address: aqeel.sahi@unisq.edu.au (A. Sahi).

https://doi.org/10.1016/j.ibmed.2025.100253
Received 16 December 2024; Received in revised form 16 April 2025; Accepted 21 April 2025
Available online 23 April 2025

feature selection. Aggarwal et al. (2022) found the most critical parameters for diagnosing PCOS. In that study, a design of Experiments (DOE) was used to minimise the number of diagnostic parameters while improving accuracy. Another study by Aggarwal et al. (2021) was conducted to diagnose PCOS. In that study, several machine learning algorithms were tested, and the authors found that 12 key features were the most significant for diagnosing PCOS. Zhang et al. (2024) combined gene expression analysis, machine learning, and network biology to diagnose PCOS. Aggarwal et al. (2023) thoroughly investigated the most essential feature for PCOS diagnosis. In that study, features like heart diseases, obesity, diabetes, and high blood pressure were found to be vital for diagnosing PCOS. Six classification models were tested and used in that study. Liu et al. [9] applied an artificial neural network model with a support vector regression (SVR) to detect PCOS. A clinical cohort composed of 1365 women was utilised in that study. Danaei et al., [10] tested several machine learning algorithms and unique feature selection algorithms to detect PCOS from clinical data. Their results showed that the feature selection algorithms improved the performance of all classifiers. Baweja et al. [11] suggested a neural network model based on rudimentary features for POCS detection. Roy et al. [12] investigated patient's personal information such as age, metabolic, biochemical factors, and marital status. Three classification models including SVM, decision tree, and Naïve Bayes were used to classify the extracted features into healthy and unhealthy subjects. Vishwakarma et al. [13] used a CNN model which was trained on 152 subjects, then it was validated and tested on 33 subjects and 32 subjects respectively to predict PCOS. Statistical metrics were used to assess their model. Bhat et al., [14] examined several classification modes in PCOS detection. They found that a linear discriminant classifier gave a superior performance compared with other machine learning models. Neto et al., [15] made a comparison among different machine learning algorithms namely, SVM, NN, RF, LR, and Naïve Bayes. They found that RF provides the best performance compared with other models. Mehrotra et al., [16] applied a *t*-test, Bayesian Classifier, and logistic regression to detect PCOS from patient's medical records. Xie et al., [17] combined CNN and a random forest model to detect PCOS. A gene ontology analysis was conducted in that study.

The discussions in the studies have highlighted a crucial observation: not all clinical data have been thoroughly examined for PCOS detection. Consequently, there is a demand for the creation of a new model that can discern the most effective features for PCOS detection while minimizing complexity and processing time. Many recent studies have tried to diagnose PCOS from ultrasound scans. For example, Rachana et al. [18] proposed a machine learning based ultrasound image model. Several image segmentation and feature extraction techniques were tested in that study. Fruh et al. [19] investigated a study on electronic medical records data of 5492 women. In that study, ultrasound images were analysed using machine-learning text algorithms. Hosain et al. [20] designed a convolutional neural network-based model to identify PCOS. In that study, ultrasound images from healthy and unhealthy subjects were used to assess their model. Panicker et al., [21] suggested CNN for detecting PCOS. The proposed model classified ultrasound images into the PCOS and non-PCOS classes. Chitra et al. [22] proposed a hybrid transfer model that combined AlexNet, Inception V3, ResNet-50, and VGG-16. Mahajan et al. [23] suggested the YOLO (You Only Look Once) model to classify ultrasound images into PCOS images or non-PCOS images. According to the above references, detection of PCOS from ultrasound images requires high-quality images, and it's more expensive.

To improve PCOS detection, a robust detection method is proposed, and the research route that contains three innovative highlights is described as follows. 1) a feature selection model named KM-GA that integrates the K-mean model and genetic algorithm is proposed. 2) a bootstrap aggregating ensembles (RSBE) model is designed to classify the selected features into healthy and unhealthy subjects. 3) Several individual and ensemble models are tested and compare their results with the proposed RSBE model.

The key scientific contribution of this paper is the provision of time series feature selection and classification methods to reduce the cost associated with PCOS detection systems and enhance the system's robustness and suitability. The remainder of this paper is organised as follows. Section 2 PCOS dataset is explained. Section 3 introduces the proposed RF ensemble-based K-mean cluster methodology framework. In Section 4, experimental results are presented and discussed. Finally, the conclusions are presented in Section 5.

2. PCOS time series dataset

This study uses a publicly available PCOS dataset to evaluate the proposed method, which was collected from Kaggle's Learning Repository [24]. The dataset is recorded from 541 Indian women aged 20–48, with heights ranging from 134 to 171 cm, and weights ranging from 31 to 108 kg. From each subject, a total of 41 true integer traits were recorded. The attribute names are {Age, Height, Body Mass Index (BMI), Blood Group, Pulse rate, Prolactin, Cycle, Marriage Status, pregnancy, No. of abortions, beta -Human Chorionic Gonadotropin, PCOS, weight, number of Cycle days, the level of follicle stimulating hormone, Luteinizing hormone, Hip, Waist, Waist: Hip Ratio, thyroid Stimulating Hormone, anti-Mullerian Hormone, Prolactin, Vitamin D3 Deficiency, Progesterone, Random Blood Sugar, Weight gain, hair growth, skin darkening, Hair loss, Pimples, Fast food, Reg Exercise, Systolic Blood Pressure, Diastolic Blood Pressure, Follicle number, average, Endometrium}.

3. Methodology

This study presents a robust model for diagnosing PCOS using clinical data. Fig. 1 shows the block diagram of the suggested model. First, the preprocessing phase is performed. Then, K-means clustering and a Genetic algorithm are implemented to select the most relevant features from the patient's clinical time series record. The outputs of the genetic algorithm and k-means are integrated, and different sets of features are formed and selected using arithmetic operators. The selected features are then used as inputs to the proposed RBSE ensemble classifier, as well as to a set of ensembles and individual classification models. In this study, to reduce potential bias in model assessment, the dataset was randomly divided into training and testing sets. We considered a ratio of 70 % for training and 30 % for testing. The proposed model was implemented using MATLAB, version R2020a, MathWorks Inc.

3.1. Data preprocessing

Medical datasets often suffer from missing values and outliers due to network loss, device failure, irregular time recording, and other factors. Several machine learning models are sensitive to those issues. We made Data preprocessing to fill in missing data. Many statistical methods have been developed to deal with missing data. Most methods rely on the percentage of missing data and the significance of the features that are missing. In the case of the missing data, which is between 5 % and 10 %, classic statistical methods such as max, mod, and mean work well. However, When the percentage of data missing is above 25 %, advanced methods such as hot deck are required. In this paper, we removed features with more than 25 % missing values. Features with missing values, which are less than 25 %, are considered. We removed columns that contain several null values. As a result, the dataset is converted into a matrix. The dimension of the resulting matrix was 538X42 where 538 refers to the number of samples and 42 indicates the number of features which are employed to diagnose PCOS.

3.2. Feature selection

This paper integrates the genetic algorithm and K-means to select the most powerful features for detecting PCOS. The outputs of the genetic

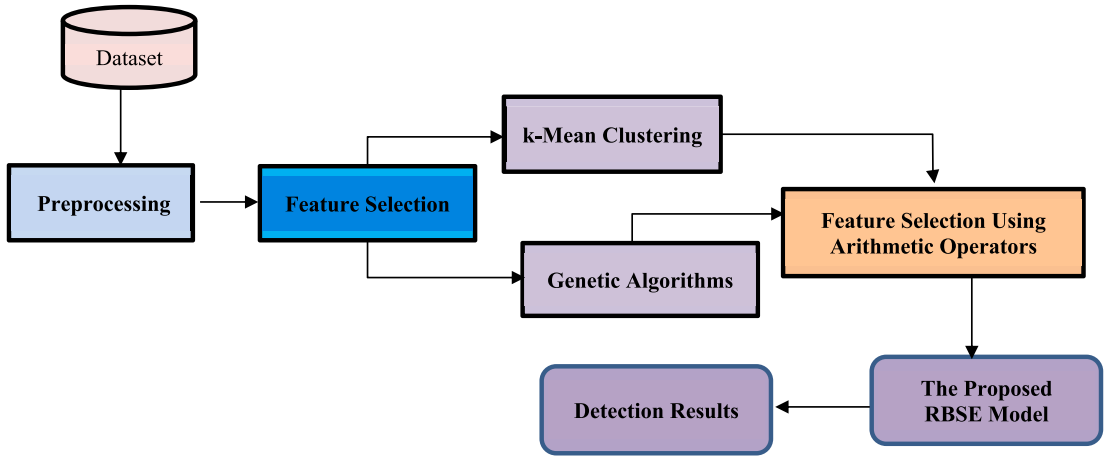

Fig. 1. Block diagram of the proposed method.

algorithm and K-means are combined based on arithmetic operators. The following section gives details regarding the genetic algorithm and K-means.

3.2.1. *Genetic Algorithms*

We employed the Genetic Algorithm (GA) to identify the most significant features from medical records. Fig. 2 describes feature selection using GA. GA generates a random population, which is used in the evaluation phase based on a fitness function. The elite's children are automatically propelled to the next generation, while the children who remained in the current society were allowed to pass genetically through the function of crossing over and mutation to form a new generation [25]. There are 41 features of PCOS in each patient's record in this dataset. The parties are either "infected" or "uninfected".

Suppose we have a set of features and need to identify the most powerful ones. A binary vector [1, 0, 0, 1, 1, …] is created where 0 refers to rejected feature, and 1 denoted to selected feature. The vector is represented as "individual", and each vector value is named a "gene". The genes are randomly chosen from {0, 1. In Fig. 3, the number of genes is N = 12, and the population size is 8. The objective function is used to evaluate everyone. In this stage, the individuals with the best objective values are selected, while individuals with the worst values are discarded. Then, a gene pool is created using crossover and mutation, as shown in Fig. 3.

Table 1 lists the parameters of GA. As previously discussed, selecting the optimal number of chromosomes is essential in the evolutionary computation phase [25,26,27]. The literature contains a variety of findings regarding the appropriate population size [28,29]. Researchers typically argue that a "small" population size could cause a poor solution [30,31] while a large" population requires a high computational time to find a solution. To define a subset feature, the trapping function that

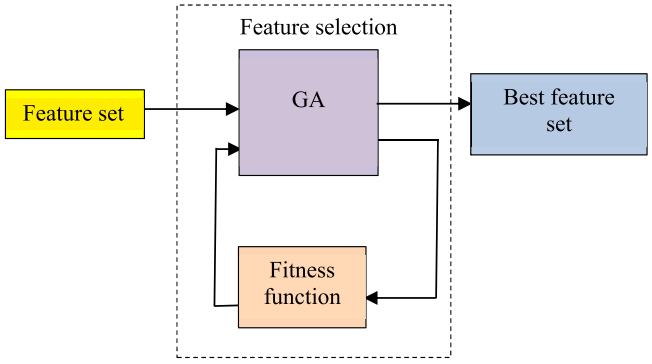

Fig. 2. Feature selection using GA.

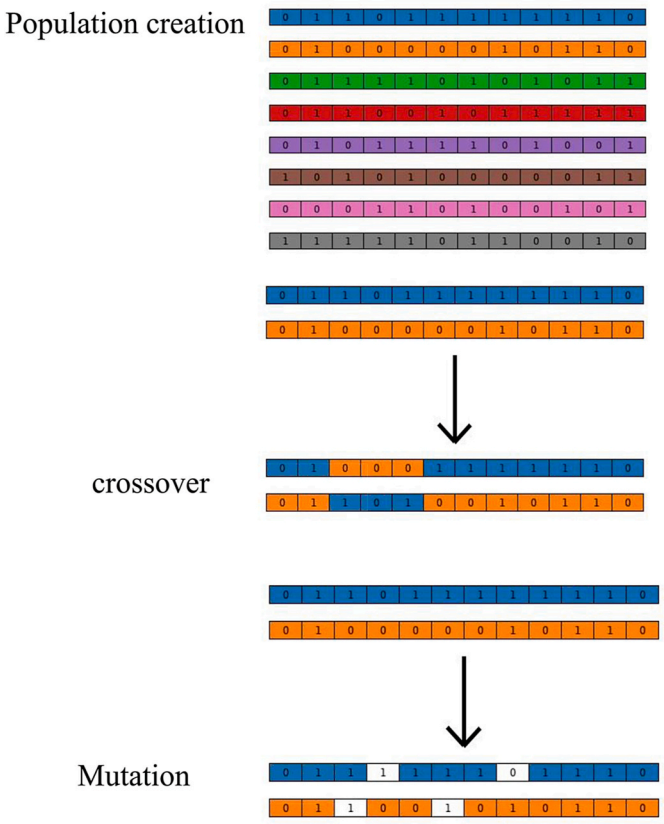

Fig. 3. The main process of GA.

evaluates the suitability of each subset feature must be defined. We adopted Oluleye's as a fitness function [26]. The distance between the training sample and the testing sample is calculated using the KNN algorithm. Individuals are assessed according to the KNN- error. All individuals with high physical fitness have a top priority to survive in the next generation. GA reduces the error rate and selects the individual with the best fitness error. This can reduce the number of features. Predictions examine the entire training sample to identify the K most similar instances, generating a new data point. Various population sizes were tested in this work to find the optimal size.

3.2.2. *K-means clustering technique*

In classification applications, the quality of the classification result is

Table 1
List of GA parameters.

GAs Parameter	Value
Number Feature	41
Population size	50,100
Genome length	41
Population type	BitstringS
Fitness Function	KNN-Based Classification Error
Number of generations	100,150
Crossover	Arithmetic crossover
Crossover Probability	0.8
Mutation	Uniform
Mutation Probability	0.1
Selection Scheme	Tournament of Size 2
EliteCount	2

heavily influenced by the features chosen. Noisy and repetitive features are removed during the feature selection process, while informative features are retained [32]. In this paper. The k-means algorithm is adopted as a feature selection model. K-Means is used to partition the dataset into two clusters. The best accuracy was achieved using a K-Means cluster with 10 replicates and k = 2, as shown in Table 2. Algorithm 1 describes the K-Means algorithm.

Algorithm 1.

Input: PCOS dataset
Output: feature set.
- Choose the number cluster.
- Mix the feature set first to initialize the centroids, then arbitrarily choose the K feature for the centroids.
- repeat step 2, until no change is detected in the middle points.
- We assigned S_i features to the closest cluster identified by the sum of the city block distance:

$$d(x,c)=\sum_{j=1}^{p}\left|x_j-c_j\right| \quad (1)$$

Recalculate the center pointer N_k for each cluster, to reflect the new tasks.

$$N_k=\frac{\sum_{i=0}^{n}x_i}{n} \quad (2)$$

3.2.3. Random subspace with bootstrap aggregating ensembles (RSBE)

Several studies based on machine learning have demonstrated that combining the outputs of multiple classifiers reduces generalisation error [33–35]. Ensemble methods are a very effective way because they combine different types of classifiers with distinctive "inductive biases" [33–42]. Indeed, such diversity used by ensemble methods can effectively reduce variance error while increasing bias error. In this work study, we proposed a method based on a group of base classifiers to detect PCOS using random subspace with bootstrap aggregating ensembles (RSBE).

We suggest an ensemble model that generates new learning sets using random subspaces and bagging [24,42–51]. In this model, we updated the training set based on two ways. First, we modified the training set by adopting bootstrap replicates $S^i=\left(X_1^i,X_2^i,\ldots X_{n,}^i\right)$ of the training set $S=(X_1,X_2,\ldots X_n)$. Then, we modified the feature space.

Suppose $X_j^i=(\mathrm{j}=1,2,\ldots,\mathrm{n};\mathrm{i}=1,2,\ldots\mathrm{B})$ of a bootstrap replicate $S^i=\left(X_1^i,X_2^i,\ldots X_{n,}^i\right)$ represented by a p-dimensional vector $X_j^i=\left(X_{j1}^i,X_{j2}^i,\ldots X_{jp,}^i\right)$. We arbitrarily chose $p^*<p$ attributes from every bootstrap replicate X^i. As a result, we obtained a p^* dimensional random subspace from the original p-dimensional feature space. The modified training set is represented as $\sim S^i=\left(\sim X_1^i,\sim X_2^i,\ldots\sim X_{n,}^i\right)$ which includes p*-dimensional training $\sim X_j^i=\left(\sim X_{j1}^i,\sim X_{j2}^i,\ldots\sim X_{jp}^i\right)$ $(\mathrm{j}=1,2,\ldots,\mathrm{n})$. The p^* components X_j^i $(\mathrm{k}=1,2,\ldots,\mathrm{p}^*)$ are randomly chosen from p components by integrating bagging and random subspaces X_{jk}^i $(\mathrm{j}=1,2,\ldots,\mathrm{p})$ of the training vector X_j^i (the selection is the same for each training vector). One then constructs base-level classifiers in the random subspaces $\sim S^i$ (of the same size), $\mathrm{i}=1,2,\ldots\mathrm{B}$ and combines them with a voting scheme in the final prediction rule. We name this algorithm Random Subspace with Bootstrap Aggregating Ensembles (R*SBE*). Table 3 lists the models used to form the RSBE.

Table 2
K-means parameters.

Parameter	Value
No. of clusters	2
Replicates	10
Distance type	City block

Table 3
Parameters of all classifiers.

Classifiers	Parameters
Linear SVM	Kernel Function = Linear
gaussian SVM	Kernel Function = Linear
KNN	Distance Function = Euclidean, K = 1,3,5
Decision Tree	Default Parameters
Naïve Bayes	Default Parameters

As shown in Fig. 4, seven algorithms were implemented to form the classification set i.e., support vector machine, Decision Tree, Nave Bayes (NB), and KNN. The weights are then assigned and calculated for each of the classifier algorithms based on their performance. Each classifier algorithm's weight is determined using the error rate as a criterion. This means that when a workbook's error rate is low, it is given a high weight. The weight β for each classification algorithm is calculated using the following equation:

$$\beta=log\frac{1-error}{error(L)} \quad (3)$$

where L is a classification algorithm.

We considered the following steps in the classification phase.

- We calculated the error rate for each individual model during the training phase.
- The error rate was calculated as follows:
  - o Let the error rates of Linear SVM, Gaussian SVM, KNN3, Naïve Bayes, Decision Tree, KNN1 = 0.16, 0.25, 0.30, 0.27 0.32, 0.12, 0.20.
  - o Using Eq.3 the models obtained the following weights: Linear SVM = 0.72, Gaussian SVM = 0.48, KNN3 = 0.37, Naïve Bayes = 0.43 and KNN5 = 0.33, KNN1 = 0.87, Decision Tree = 0.60.
  - o Assume each model classifier identifies a targeted PCOS segment as follows: Linear SVM = C1, Gaussian SVM = C2, KNN3 = C2, Naïve Bayes = C2 and KNN5 = C1, KNN1 = C1, Decision Tree = C2.
  - o Based on the ensemble model in Fig. 4, the weighted vote was calculated as Class (C1): Linear SVM + KNN5 + KNN1→ 0.72 + 0.33+ 0.87 = 1.92 Class (C2) = Gaussian SVM + KNN3+ Naïve Bayes + Decision Tree → 0.48 + 0.37+0.43 + 0.60 = 1.88.
  - o As a result, the class (C1) obtained a higher value than class (C2). The ensemble classifier considered the targeted segment as the PCOS segment.

3.2.4. Performance evaluation

Several metrics are used for performance evaluation. In this study, recall, F-measure, accuracy, sensitivity, and specificity are employed to the proposed method [51–58].

- $Recall=\frac{TP}{TP+FN}$

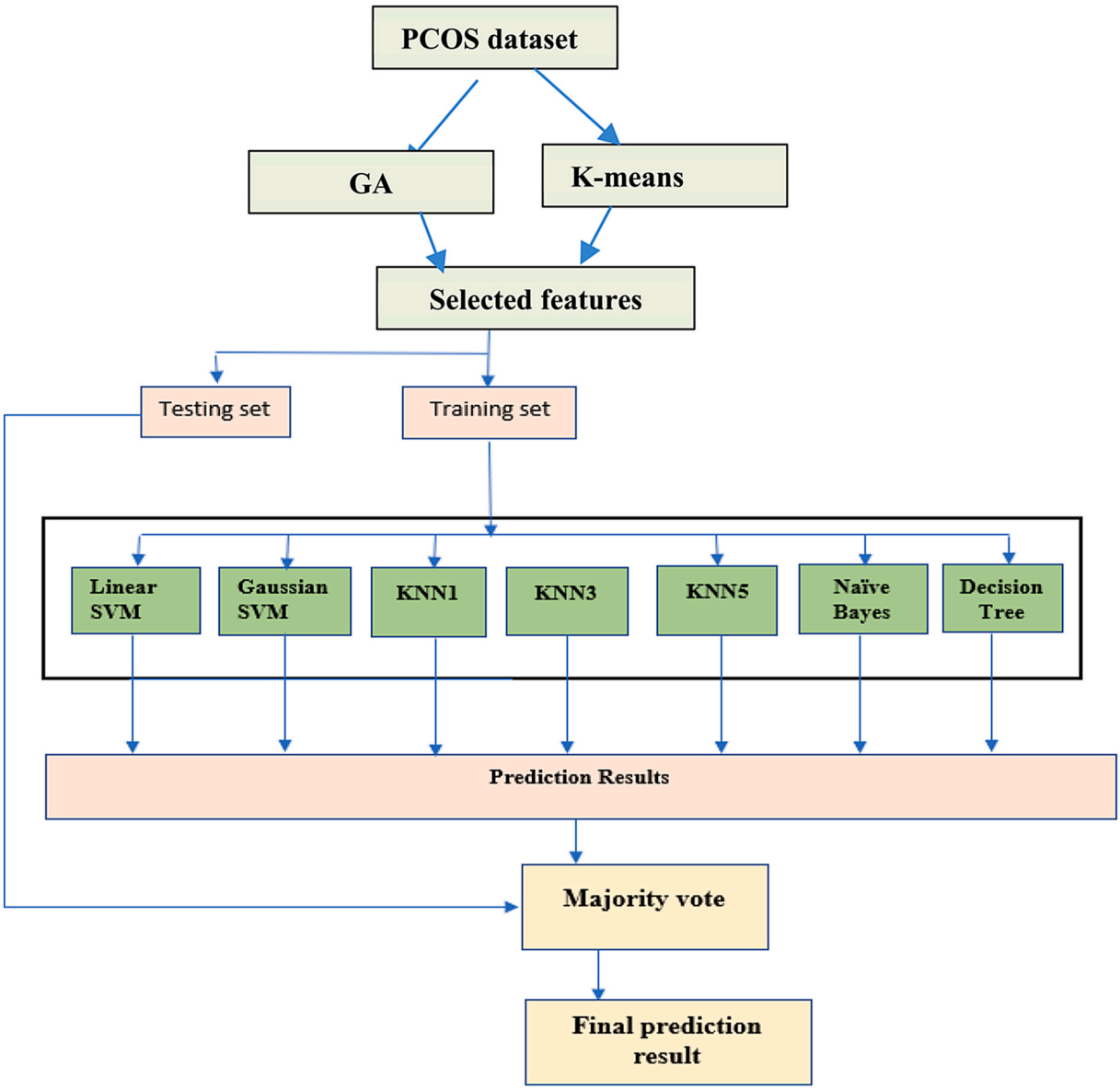

Fig. 4. The proposed ensemble model.

- $Precision = \frac{TP}{TP+FP}$
- $F\text{-}measure = 2 \times \frac{Precision*Recall}{Precision+Recall}$
- $Specificity = \frac{TN}{TN+FP}$
- $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$.

Where TP refers to the person with no PCOS symptoms, TN denotes to the PCOS patient correctly recognised as a PCOS patient. FN denotes the patient with PCOS and is classified as a healthy person. FP refers to a healthy patient while predicted as a PCOS patient.

4. Experimental results

In this paper, data from 538 patients were collected, of which a total of 421 subjects were healthy, while 177 subjects were identified as PCOS subjects. The data was preprocessed to remove unwanted columns, and then we standardised the data. As mentioned before, two feature selection models were integrated in this study. K-means and Genetic Algorithm were employed to select the optimal feature set. As a result, a total of six features were selected using the K-Means algorithm, and a total of ten features were selected by the Genetic Algorithm. The selected features were sent to the proposed model RSBE. An accuracy of 95.68 was obtained when the RSBE was combined with k-Means, and an accuracy of 91.98 was gained when the RSBE was combined with the Genetic Algorithm. Table 4 reports the classification accuracy of the proposed model.

4.1. Diagnostic results based on genetic algorithm

In this experiment, it was observed that the proposed ensemble

Table 4
The classification results based on two feature selection models.

The proposed model	Sensitivity	F-Measure	Precision	Specificity	Accuracy
RSBE based on k-Means	94.12	93.20	92.31	96.40	95.68
RSBE is based on a Genetic Algorithm	76.36	86.60	87.4	85.5	89.98

model RSBE coupled with the Genetic Algorithm. The highest accuracy rate was scored using GA when the population number was set to 50,100 and k = 3 for the Fitness Function (KNN). The number of features extracted by the Genetic Algorithm was 7. To find the best solution, different values of *k* were tested. The fitness function was chosen carefully to minimise the classification error. A total of 7 chromosomes were selected from the total of 41, as shown in Fig. 5. Fig. 5 lists the best value for fitness using KNN. In addition, Fig. 6 shows the best and worst scores of the fitness function. The classification error for the PCOS was 0.00185874, 0.00221516. Table 5 shows the classification results based on the selected features named Fast Food, hair growth, Cycle length (days), RR, VD3, Pimples, Follicle No. (R)).

Table 5 presents the results of PCOS detection based on ensemble methods using a Generic Algorithm as feature selection. The proposed RSBE achieved the highest specificity and precision compared with the other models. The performance of all ensemble models was degraded with a genetic algorithm. Our findings showed that the low performance resulted from some noisy features that were selected by Genetic Algorithm. The basic ensemble had a better performance, it achieved 91.98 %. The Boosted Ensemble recorded the lowest performance with an f-score 86.67, sensitivity, 84.76.

4.2. Diagnose results based ON K-MEANS

With k-means, the proposed model achieved the best accuracy when K-means parameters were set to Replicates = 10 and $k = 2$. More details regarding the results obtained are in the next section. Table 6 shows the detection rate based on the selected features. The six selected features were ranked based on the accuracy rate. The follicle feature was ranked the top feature, gaining the highest detection. However, weight gain recorded the lowest detection rate. The detection results of integrating K-means with RSBE are listed in Table 7. In this experiment, the proposed ensemble model was compared with other ensemble approaches. The selected features using K-Means were sent to the RSBE as well as to the ensemble methods. Table 7 presents the comparison results among the proposed model RSBE, and other ensemble classifiers. The RSBE scored a high accuracy compared with other ensemble models. However, the random subspace ensemble recorded a higher specificity than the RSBE accuracy of 95.68 %. The bagging ensemble achieved a classification accuracy of 95.06 %, Prec. of 89.58 % and sensitivity of 93.48 %. Our results showed that the bagging ensemble produced a higher number of false positives and a lower number of false negatives.

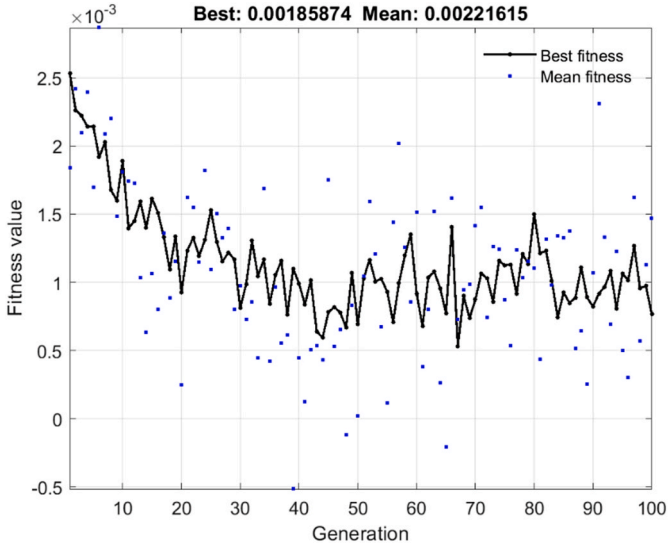

Fig. 5. The optimal fitness value.

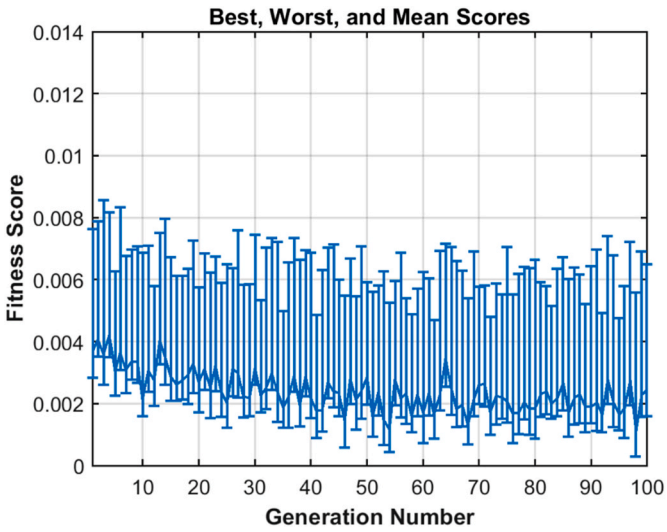

Fig. 6. The highest and lowest scores of the fitness function.

4.3. Integrating K-MEANS and genetic algorithm

To improve diagnose of PCOS, we integrated the output of K-Means and Genetic Algorithm. We used a mathematical operator to find out the best combination of features. Table 8 reports the accuracy of integrating k-Means with Genetic Algorithms. We can notice that the classification accuracy was improved when the selected features by k-Means with Genetic Algorithm using ∪ operator.

4.4. RSBE performance evaluation with individual classification models

The proposed RSBE model was compared with several individual classification models. Firstly, the comparisons were conducted based on the k-mean feature selection model. In this experiment, the K-Means were used as a feature selection model, and the selected features were sent to several individual classifiers as well as to the RSBE model. We found that the K-Mean model showed a high performance with all classification models. The main cause is that K-Mean was capable to select the most influential attributes as well as eliminating the irrelevant feature. The classification results for different classifiers are presented in Table 9. As shown in Table 9, the naive bayes algorithm and Linear SVM algorithm scored very close results with an accuracy of 89.51 %. The Naive Bayes algorithm scored an increase in F-Measure, Precision, sensitivity, and Specificity. However, the proposed model RSBE recorded the highest accuracy compared with all individual classifiers.

Second, in this experiment, the best features selected by the genetic algorithm were fed to individual classifiers. Table 10 shows that the accuracy of weak individual classifiers was lower than that of the proposed ensemble model. For individual classifiers, it can be noticed that the KNN3 and Linear SVM algorithms achieved the highest accuracy of 89.51 % and 88.72 %, respectively. However, the KNN1 recorded the lowest accuracy among the individual classification models. In addition, compared to Table 10, we can observe that K-Means performed very well with individual classifiers compared with the Genetic Algorithm. Further evaluation was made using a 10-fold cross-validation metric. The results in Fig. 7 confirmed our findings in Tables 9 and 10.

According to Fig. 8, which depicts SHAP values for the detection of Polycystic Ovary Syndrome (PCOS), we can identify the relative importance of various features for both the normal group (represented by blue bars) and the diagnosis group (shown in orange). The chart illustrates how specific biological and symptomatic characteristics are weighed differently in predicting PCOS, with Height, Progesterone, and Pimples standing out as key indicators for diagnosis.

We applied the Wilcoxon Signed-Rank Test to evaluate the results

Table 5
Typical Performance Test Results for Ensemble Methods Using ga.

Ensemble techniques	Sen.	F-Measure	Prec.	Spe.	Accuracy	Confusion Matrix			
						Predict Class			
						0	1		
Basic Ensemble	90.38	91.26	92.16	96.36	94.44	106	4	0	True Class
						5	47	1	
Bagged Ensemble	93.88	89.32	85.19	92.92	93.21	105	8	0	
						3	46	1	
Boosted Ensemble	84.78	86.67	88.64	95.69	92.36	111	5	0	
						7	39	1	
Random Subspace Ensemble	91.30	87.50	84.00	93.10	92.59	108	8	0	
						4	42	1	
RSBE	76.36	86.60	1	1	91.98	107		0	
						13	42	1	

Table 6
Classification results using the K-Means model.

Features	Detection rate
Follicle	78.7
SK	77.6
FN	77.5
HG	77.1
CL	74.3
WG	74.2

Skin darkening (SK), Follicle No. (R) (FN), Hair growth level (HG), Cycle length (CL), Weight gain (WG).

obtained. The Wilcoxon Signed-Rank Test is considered one of the most effective non-parametric tests that compare the performance of two models' accuracies. Table 11 displays the observed results. From the results, it can be noticed that the proposed model outperformed the state-of-the-art models.

4.5. Comparison with state-of-the-art methods

In this section, the proposed model was compared with several previous methods. All studies were tested with the same dataset. Kanvinde et al., [38] designed a Bootstrap ensemble model to detect PCOS. Bharati et al., [5] detected PCOS based on a filter-based univariate feature selection method. In that study, they considered 10 features.

Table 8
Classification results using K-means integrated with genetic algorithm.

	Accuracy
K-Means ∩ Genetic Algorithm	86 %
K-Means ∪G enetic Algorithm	99 %

Table 7
Typical performance test results for ensemble methods using K-mean Cluster.

Ensemble techniques	Sensitivity	F-Measure	Prec.	Specificity	Accuracy	Confusion Matrix			
						Predict Class			
						0	1		
Basic Ensemble	90.00	89.12	88.24	94.84	93.21	108	6	0	True Class
						5	45	1	
Bagged Ensemble	93.48	91.49	89.58	95.69	95.06	108	5	0	
						5	15	1	
Boosted Ensemble	82.98	84.78	86.67	94.78	91.36	109	6	0	
						116	39	1	
Random Subspace Ensemble	86.67	91.76	97.50	99.15	95.68	116	1	0	
						6	39	1	
RSBE	94.12	93.20	92.31	96.40	95.69	107	4	0	
						3	48	1	

Table 9
Performance comparisons based on individual classifiers and K-means feature Selection.

Classifier	Sen.	F-Measure	Prec.	Spec.	Acc.	Confusion Matrix			
						Predict results			
						0	1		
Linear SVM model	84.8	84.2	83.4	91.8	89.6	100	9	0	True Class
						8	45	1	
Gaussian SVM model	83.4	81.9	80.4	89.9	87.7	97	11	0	
						9	45	1	
Decision Tree algorithm	78.8	80.5	82.5	92.1	87.8	101	9	0	
						11	41	1	
Naïve Bayes model	85.9	86.8	87.6	91.3	89.7	90	8	0	
						9	29	1	
KNN1 model	84.6	87.7	73.5	90.5	88.5	111	12	0	
						6	33	1	
KNN3 model	76.8	80.21	83.31	92.8	87.7	102	8	0	
						12	40	1	
KNN5 model	74.6	78.9	83.8	92.62	86.5	99	8	0	
						14	41	1	
The proposed RSBE	94.12	93.3	92.4	96.5	95.8				

Table 10
Performance comparisons based on individual classifiers and genetic algorithm feature Selection.

Models	Sen.	F-Measure	Prec.	Spec.	Acc	Confusion Matrix			
						Predict Class			
						0	1		
Linear SVM model	78.2	80.5	83.1	92.9	88.3	99	12	0	True Class
						12	39	1	
Gaussian SVM model	63.49	75.47	93.03	96.97	83.95	104	8	0	
						11	39	1	
Decision Tree model	76.47	76.47	76.47	89.19	85.19	96	3	0	
						23	40	1	
Naïve Bayes model	82.1	78.9	76.1	88.4	86.5	91	14	0	
						14	43	1	
KNN1 model	75.5	75.6	75.5	86.8	82.6	98	8	0	
						9	99	1	
KNN3 model	84.5	85.3	86.1	92.4	89.6	101	8	0	
						13	40	1	
KNN5 model	75.5	79.3	83.4	88.5	87.1	99	13	0	
						9	4	1	

Compared with our results, we used lower features to detect PCOS. Munjal et al., [39] proposed a genetic algorithm to detect PCOS. In that study, random forest, and decision trees were used to classify the features into healthy and unhealthy subjects. Tanwani et al. [40] suggested a method based on two machine learning algorithms, k-nearest and logistic regression. 20 features were used in that study. Table 12 shows the comparison of our proposed method with state-of-the-art.

5. Conclusion

Early detection of PCOS is crucial for prompt patient treatment. An automated system that relies on clinical and metabolic parameters could

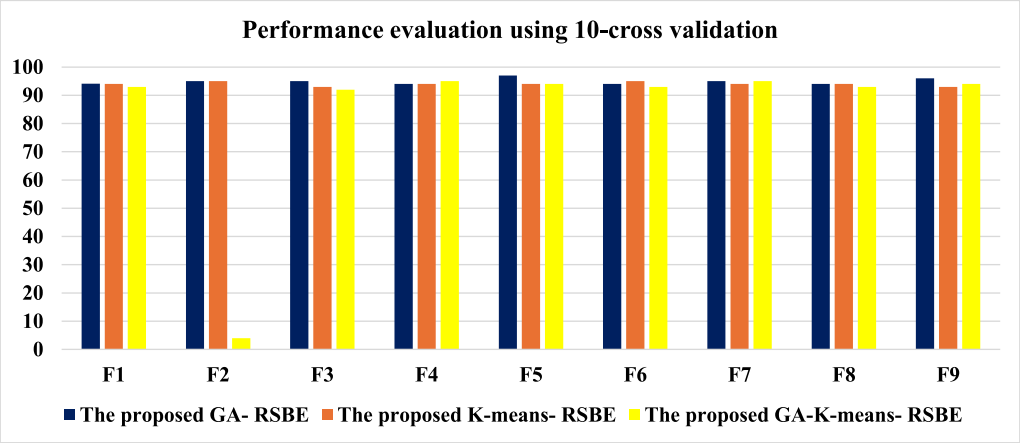

Fig. 7. Evaluation using 10-fold cross-validation.

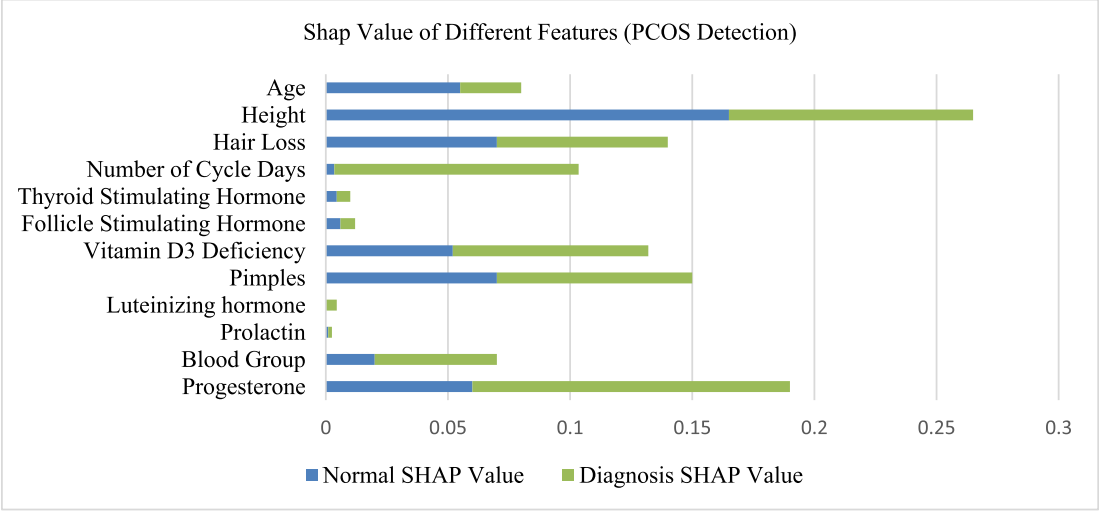

Fig. 8. Features importance using SHAP value.

Table 11
Wilcoxon test for PERFORMANCE comparisons.

Model	Wilcoxon p-value	Results
RSBE vs Linear SVM model v	0.0021	RSBE significantly better
RSBE vs Gaussian SVM model	0.0033	RSBE significantly better
RSBE vs Decision Tree algorithm	0.0012	RSBE significantly better
RSBE vs Naïve Bayes model	0.0032	RSBE significantly better
RSBE vs KNN1 model	0.0024	RSBE significantly better
RSBE vs KNN3 model	0.0034	RSBE significantly better
RSBE vs KNN5 model	0.0032	RSBE significantly better

Table 12
Comparison with the state-of-the-art.

Authors	Accuracy
Kanvinde, N. et al. [44]	92.00 %
Bharati,et al. [5]	91.01 %
Munjal,et al. [24]	88.00 %
Tanwani, N. et al. [46]	92.00 %
Proposed method	98 %

be a valuable tool for PCOS identification. While numerous automated detection systems have been proposed in the literature, many of them rely on deep learning techniques, which demand large datasets for accurate performance. However, due to data availability constraints, numerous studies fail to meet this requirement, resulting in reduced accuracy, sensitivity, or increased computation time when employing individual machine learning algorithms. In our research, we took a different approach, employing a variety of ensemble methods and machine-learning algorithms for PCOS detection. Our dataset, sourced from the Kaggle repository, comprises 541 samples. Specifically, we evaluated the performance of six well-established and capable machine learning approaches: Stacked Ensemble, Random Subspace, Boosted Ensemble, Bagged Ensemble, Ensemble Learning, and the RF model.

The experimental findings reveal that our proposed method, which combines a subspace ensemble model and RSBE learning model, alongside features extracted using the K-Means algorithm delivers superior classification performance in predicting PCOS cases across most of the scenarios considered.

CRediT authorship contribution statement

Najlaa Faris: Writing – original draft, Formal analysis, Data curation, Conceptualization. **Aqeel Sahi:** Validation, Formal analysis, Data curation. **Mohammed Diykh:** Formal analysis, Data curation, Conceptualization. **Shahab Abdulla:** Writing – review & editing. **Siuly Siuly:** Writing – review & editing.

Data availability

The datasets used for testing this study were drawn from open-access databases [42].

Ethics statement

No ethical approval was required for this study.

Declaration of competing interest

The authors declare that there are no financial interests to influence the study reported in this paper.

References

[1] Zhang X, et al. Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening. Mol Cell Endocrinol 2021;523(December 2020):111139. https://doi.org/10.1016/j.mce.2020.111139.

[2] Mohammad Ali NFH, Megat Hanafiah MAK, Saleh SH, Mohd Ali MT, Ibrahim S. A review of biomass-based natural coagulants for water pollution remediation: impact of properties and coagulation operational parameters. AUIQ Complementary Biological System 2024;1(2):31–45.

[3] Kusrini E, Hashim F, Aziz A, Hassim MFN, Usman A, Wilson LD, Negim ES, Prasetyo AB. Potential of critical mineral and cytotoxicity activity of gadolinium complex as anti-amoebic agent by viability studies and flow cytometry techniques. AUIQ Complementary Biological System 2024;1(2):1–10.

[4] Maheswari K, Baranidharan T, Karthik S, Sumathi T. Modelling of F3I based feature selection approach for PCOS classification and prediction. J Ambient Intell Hum Comput 2021;12(1):1349–62. https://doi.org/10.1007/s12652-020-02199-1.

[5] Agha H, Sanaalla AB, Abdulsamad MAS, Mohammad GAR, Allaq AAA. Impact of some parameters on the survival and proliferation of foodborne pathogens: Escherichia coli, Bacillus subtilis, Staphylococcus aureus, and Streptococcus pyogenes. AUIQ Complementary Biological System 2024;1(1):70–6.

[6] Aghaa HM, Musa SA, Hapiz A, Wu R, Al-Essa K, Saleh AM, Reghioua A. Biocomposite adsorbent of grafted chitosan-benzaldehyde/lactobacillus casei bacteria for removal of acid red 88 dye: box-benken design optimization and mechanism approach. AUIQ Complementary Biological System 2025;2(1):1–14.

[7] Bhosale S, Joshi L, Shivsharan A. "PCOS (POLYCYSTIC OVARIAN SYNDROME) DETECTION USING," no. 01. 2022. p. 195–200.

[8] Soni P, Vashisht S. Exploration on polycystic ovarian syndrome and data mining techniques. In: Proc. 3rd int. Conf. Commun. Electron. Syst. ICCES 2018; 2018. p. 816–20. https://doi.org/10.1109/CESYS.2018.8724087. Icces.

[9] Liu L, Shen F, Liang H, Yang Z, Yang J, Chen J. Machine learning-based modeling of ovarian response and the quantitative evaluation of comprehensive impact features. 2022.

[10] Danaei Mehr H, Polat H. Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. Health Technol 2022;12(1):137–50.

[11] Baweja AK, Gupta M. Neural network-based method to predict PCOS in women. In: Proceedings of international conference on recent trends in computing; 2022. p. 227–36.

[12] Roy DG, Alvi PA. Artificial intelligence in diagnosis of polycystic ovarian syndrome. In: Contemporary issues in communication, cloud and big data analytics; 2022. p. 453–63.

[13] Vishwakarma V, Chethan S, Datla MT, Aqib MM, Roy S, Thasni T. Prediction of severity of polycystic ovarian syndrome using artificial neural networks. In: International conference on image processing and capsule networks; 2021. p. 589–98.

[14] Bhat SA. Detection of polycystic ovary syndrome using machine learning algorithms, 2020; 2021. p. 1124–7.

[15] Neto C, Silva M, Fernandes M, Ferreira D, Machado J. Prediction models for Polycystic Ovary Syndrome using data mining. In: International conference on advances in digital science; 2021. p. 210–21.

[16] Mehrotra P, Chatterjee J, Chakraborty C, Ghoshdastidar B, Ghoshdastidar S. Automated screening of polycystic ovary syndrome using machine learning techniques. In: Proc. - 2011 annu. IEEE India conf. Eng. Sustain. Solut. INDICON-2011, no. 1; 2011. https://doi.org/10.1109/INDCON.2011.6139331.

[17] Xie NN, Wang FF, Zhou J, Liu C, Qu F. Establishment and analysis of a combined diagnostic model of polycystic ovary syndrome with random forest and artificial neural network. BioMed Res Int 2020;2020. https://doi.org/10.1155/2020/2613091.

[18] Rachana B, Priyanka T, Sahana KN, Supritha TR, Parameshachari BD, Sunitha R. Detection of polycystic ovarian syndrome using follicle recognition technique. Glob. Transitions Proc. 2021;2(2):304–8. https://doi.org/10.1016/j.gltp.2021.08.010.

[19] Fruh V, Cheng JJ, Aschengrau A, Mahalingaiah S, Lane KJ. Fine particulate matter and polycystic ovarian morphology. Environ. Heal. 2022;21(1):1–8. https://doi.org/10.1186/s12940-022-00835-1.

[20] V. Krishnaveni, C. Deepa, R. S. Cindhu, and K. G. Santhiya, "An ANN based Screener for the early diagnose of Polycystic Ovarian Syndrome in adolescent and young women," vol. 13, no. 2, pp. 84–95.

[21] Thomas N. A literature inspection on polycystic ovarian morphology in women using data mining methodologies. Int J Adv Res Comput Sci 2018;9(1):547–51. https://doi.org/10.26483/ijarcs.v9i1.5393.

[22] Lele Priyanka R, Thakare Anuradha D. Comparative analysis of classifiers for polycystic ovary syndrome detection using various statistical measures. Int J Eng Res 2020;V9(3):410–2. https://doi.org/10.17577/ijertv9is030404.

[23] Denny A, Raj A, Ashok A, Ram CM, George R. I-HOPE: detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques. In: Ieee Reg. 10 annu. Int. Conf. Proceedings/TENCON, 2019-Octob; 2019. p. 673–8. https://doi.org/10.1109/TENCON.2019.8929674.

[24] Sharkey AJC. Linear and order statistics combiners for pattern classification. In: Combining artificial neural nets. Springer; 1999. p. 127–61.

[25] Babatunde OH, Armstrong L, Leng J, Diepeveen D. Zernike moments and genetic algorithm: tutorial and application. 2014.

[26] Alander JT. On optimal population size of genetic algorithms. In: CompEuro 1992 Proceedings computer systems and software engineering; 1992. p. 65–70.

[27] Diaz-Gomez PA, Hougen DF. Initial population for genetic algorithms: a metric approach. Gem 2007:43–9.

[28] Reeves CR. Using genetic algorithms with small populations. ICGA 1993;5:90–2.

[29] Roeva O. Improvement of genetic algorithm performance for identification of cultivation process models. Adv. Top. Evol. Comput. B. Ser. Artif. Intell. Ser. 2008:34–9.

[30] Koumousis VK, Katsaras CP. A saw-tooth genetic algorithm combining the effects of variable population size and reinitialization to enhance performance. IEEE Trans Evol Comput 2006;10(1):19–28.

[31] Pelikan M, Goldberg DE, Cantú-Paz E. Bayesian optimization algorithm, population sizing, and time to convergence. Livermore, CA (United States): Lawrence Livermore National Lab.(LLNL); 2000.

[32] Izzo D, Sprague CI, Tailor DV. Machine learning and evolutionary techniques in interplanetary trajectory design. In: Modeling and optimization in space engineering. Springer; 2019. p. 191–210.

[33] Quinlan JR. Bagging, boosting, and C4. 5. Aaai/Iaai 1996;1:725–30.

[34] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach Learn 1999;36(1):105–39.

[35] Opitz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res 1999;11:169–98.

[36] Domingos P. Using partitioning to speed up specific-to-general rule induction. In: Proceedings of the AAAI-96 workshop on integrating multiple learned models; 1996. p. 29–34.

[37] Mitchell TM. Machine learning. New York: McGraw-hill; 1997.

[38] Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural Comput 1992;4(1):1–58.

[39] Alsafy I, Diykh M. Developing a robust model to predict depth of anesthesia from single channel EEG signal. Physical and Engineering Sciences in Medicine 2022;45(3):793–808.

[40] Al-Hadeethi H, Abdulla S, Diykh M, Green JH. Determinant of covariance matrix model coupled with adaboost classification algorithm for EEG seizure detection. Diagnostics 2021;12(1):74.

[41] Mohammed H, Diykh M. Improving EEG major depression disorder classification using FBSE coupled with domain adaptation method based machine learning algorithms. Biomed Signal Process Control 2023;85:104923.

[42] Abdulla S, Diykh M, Siuly S, Ali M. An intelligent model involving multi-channels spectrum patterns based features for automatic sleep stage classification. Int J Med Inf 2023;171:105001.

[43] Diykh M, Abdulla S, Oudah AY, Marhoon HA, Siuly S. A novel alcoholic EEG signals classification approach based on adaboost k-means coupled with statistical model. In: Health information science: 10th international conference, HIS 2021, Melbourne, VIC, Australia, october 25–28, 2021, proceedings. 10. Springer International Publishing; 2021. p. 82–92.

[44] Lafta R, Zhang J, Tao X, Li Y, Diykh M, Lin JCW. A structural graph-coupled advanced machine learning ensemble model for disease risk prediction in a telehealthcare environment. Big data in engineering applications. 2018. p. 363–84.

[45] Diykh M, Abdulla S, Deo RC, Siuly S, Ali M. Developing a novel hybrid method based on dispersion entropy and adaptive boosting algorithm for human activity recognition. Comput Methods Progr Biomed 2023;229:107305.

[46] Ali KM, Pazzani MJ. Error reduction through learning multiple descriptions. Mach Learn 1996;24(3):173–202.

[47] Bartlett P, Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers. Adv. Kernel methods—support vector Learn. 1999. p. 43–54.

[48] Kanvinde N, Gupta A, Joshi R. Binary classification for high dimensional data using supervised non-parametric ensemble method. http://arxiv.org/abs/2202.07779; 2022.

[49] Munjal A, Khandia R, Gautam B. A machine learning approach for selection of polycystic ovarian syndrome (pcos) attributes and comparing different classifier performance with the help of wekaand pycaret. Int J Sci Res 2020;(2277):59–63. https://doi.org/10.36106/ijsr/5416514.

[50] Tanwani N. Detecting PCOS using machine learning. Int. J. Mod. Trends Eng. Sci. 2020;(1).

[51] Chew X, Khaw KW. Polycystic Ovarian Syndrome (PCOS) classification and feature selection by machine learning techniques 2020;9(MI):65–74.

[52] www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos. last accessed on 5 May 2020.

[53] van der Ham K, Barbagallo F, van Schilfgaarde E, Lujan ME, Laven JS, Louwers YV. The additional value of ultrasound markers in the diagnosis of polycystic ovary syndrome. Fertil Steril 2024.
[54] Wang M, Zhang S, He J, Zhang T, Zhu H, Sun R, Yang N. Biochemical classification diagnosis of polycystic ovary syndrome based on serum steroid hormones. J Steroid Biochem Mol Biol 2025;245:106626.
[55] Aggarwal S, Pandey K. Early identification of PCOS with commonly known diseases: obesity, diabetes, high blood pressure and heart disease using machine learning techniques. Expert Syst Appl 2023;217:119532.
[56] Zhang W, Wu Y, Yuan Y, Wang L, Yu B, Li X, Yao Z, Liang B. Identification of key biomarkers for predicting atherosclerosis progression in polycystic ovary syndrome via bioinformatics analysis and machine learning. Comput Biol Med 2024;183:109239.
[57] Aggarwal S, Pandey K, Senior Member IEEE. Determining the representative features of polycystic ovary syndrome via Design of Experiments. Multimed Tool Appl 2022;81(20):29207–27. Aggarwal, S. and Pandey, K., 2021. An analysis of PCOS disease prediction model using machine learning classification algorithms. Recent Patents on Engineering, 15(6), pp.53–63.
[58] Aggarwal Shivani, Pandey Kavita. PCOS diagnosis with commonly known diseases using hybrid machine learning algorithms. In: 2023 6th international conference on contemporary computing and informatics (IC3I). 6. IEEE; 2023. p. 1658–62.