

OPTIMIZED FEATURE SELECTION FOR PCOS DISEASE PREDICTION

Selvaraj Santhi[#], Sundaradhas Selva Nidhyananthan

Received on April 24, 2023
Presented by B. Petrunov, Member of BAS, on June 27, 2023

Abstract

PCOS is the most predominant endocrine problem among women of reproductive age, which affects ovaries and causes irregularities in menstrual cycles, weight gain, and hirsutism [1]. The condition probably results from a mix of causes, including qualities and ecological elements. The aim of this study is to investigate the performance of Binarized Butterfly Optimization Algorithm, Binarized Grey Wolf Optimization Algorithm, Binarized Genetic Algorithm and Binarized Cuckoo Optimization Algorithm in terms of classification accuracy. Kaggle PCOS dataset is used for this work and it has 541 records and 43 attributes. The study aims to investigate the performance of four optimization algorithms in terms of classification accuracy for predicting PCOS disease. The BGWO optimization algorithm outperformed the other algorithms with 99% accuracy, while the BGA produced the most optimal subset. The study highlights the significance of feature selection in improving the performance of classification algorithms.

Key words: polycystic ovarian syndrome, feature selection, optimization algorithms, classification, accuracy

1. Introduction. Among women of conceptive age, PCOS is the most predominant endocrine problem [2]. PCOS influences ovaries and causes sporadic or missing periods, ovarian growths, extended ovaries, overabundance androgen, weight gain and hirsutism. It disrupts the regular menstrual cycle and causes

[#]Corresponding author.

DOI:10.7546/CRABS.2023.09.14

irregularities when reproductive hormones are disturbed. According to statistics from WHO, 3.4% of women worldwide have PCOS. Current developments in AI have brought encouraging results in medical diagnostics. The authors of [3] employed machine learning methods, to predict the PCOS based on Kaggle dataset. AMSY et al. [4] suggested possible features based on the importance of traits using machine learning algorithms. The main objective is to develop an effective and optimal algorithm that increases PCOS diagnostic accuracy by choosing the minimal features possible. Feature selection optimization algorithms are commonly used in machine learning applications to reduce the complexity of models, improve performance, and enhance interpretability. Four optimization algorithms with binarized representation are used in this work. In the context of feature selection, classifiers are used to analyse the applicability and its performance serves as the objective function. Feature selection optimization algorithms look for the best possible combination of features to maximize the classifier's performance. In this work, multiple classifiers (Decision tree, KNN, Gaussian and SVM) are used to ensure that the selected features are robust, and not biased towards a particular type of classifier. The remaining section of the article is organized as follows: Section 2 contains a brief literature survey on PCOS diagnosis and prediction and the proposed technique is discussed in Section 3. The outcomes and conclusion of the suggested method are covered in Sections 4 and 5.

2. Literature review. RANI et al. [2] proposed a method to control PCOS with the benefits of polyphenols. SONI and VASHISHT [5] predicted PCOS using data mining techniques. NEETHA et al. [6] detected the presence of PCOS with hybrid classification like Bayesian and Neural Network classifier. DESHPANDE and WAKANKAR [7] detected PCOS by counting the number of follicles in an ovarian ultrasound picture and this method's attained accuracy is 95%. The works [8–10] proposed different methods for detecting and predicting PCOS. ALAM et al. [11] investigated the utilization and applications of Genetic Algorithms in engineering pedagogy. TUBISHAT et al. [12] enhanced BOA algorithm for feature selection; and RAJALAXMI et al. [13] designed a system utilizing BIGWO variations to select the most significant symptoms of Parkinson's disease. The earlier works [4, 14, 15] focused on identifying potential features and employing classification methods for predicting PCOS using machine learning techniques. The existing works [3, 16–18] have explored various methods for PCOS prediction and these studies demonstrated the potential of these methods.

3. Research methodology. This proposed work helps to diagnose PCOS by analysing the symptoms of the patients. Figure 1 shows the design of the proposed framework.

3.1. Data set description and cleaning. The data is collected from ten different hospitals in Kerala, India. It consists of 541 instances and 43 attributes including target attribute. Target class labels in this dataset are 1 for women with PCOS and 0 for women without PCOS. The PCOS Kaggle dataset has to be

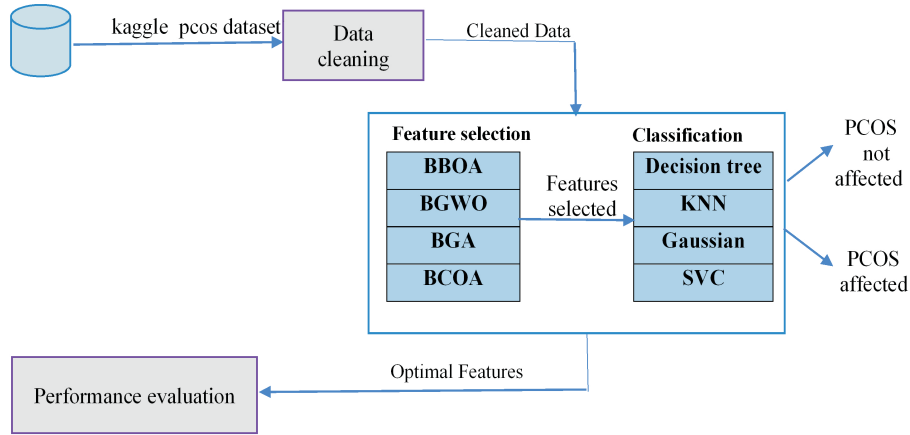


Fig. 1. Proposed system design

pre-processed, among 43 attributes; the most important 29 attributes are chosen according to gynaecologist's suggestion and those are sub-symptoms of selected symptoms. All the missing values are removed and the data is validated. Out of 29 attributes, 11 attributes are categorical and 18 attributes are numerical values.

3.2. Feature selection. 3.2.1. Binarized Butterfly Optimization Algorithm.

The BBOA is a meta heuristic algorithm that can be used to solve optimization problems in various fields. In BBOA, the search process is modelled as a collec-

Algorithm 1: BBOA

Input: Population Size PS, Input features N, Sensory modality c,

Stimulus Intensity I, max_iter t, alpha a, epoch p

Output: Optimal subset g^* , Accuracy

BEGIN

 while max_iter not reached do

 for each feature n in population N do

 compute the fragrance using Eq1

 Generate random r in [0,1]

 if $r < p$ then

 Move for best feature using $x_i^{t+1} = x_i^t + (r^2 * g^* - x_i^t) \times f_i$
 (global search)

 else

 Move randomly using $x_i^{t+1} = x_i^t + (r^2 * x_j^t - x_k^t) \times f_i$
 (local search)

 end if

 compute the Fitness function for the new solution

 end for

 end while

 return the optimal subset g^* of features

END

tion of virtual butterflies that communicate with each other using pheromone-like information. The search is directed toward better solutions using this data and Algorithm 1 describes BBOA

$$(1) \quad f = cI^a,$$

where I is level of stimulation, a is power exponent coefficient, and c is sensory modality coefficient. The algorithm's speed of convergence is significantly influenced by the parameters $a = 0.1$ and $c = 0.01$ with the help of randomly generated threshold in fitness function. There are two main phases for updating butterflies' positions: Global search and Local search phase.

3.2.2. Binarized Grey Wolf Optimization Algorithm. The algorithm mimics the wolf's hierarchical social structure, with alpha, beta, and gamma wolves serving as leaders. An initial population of wolves is utilized by the BGWO algorithm, each representing a potential solution to the optimization problem. These wolves communicate with one another and adjust their positions in accordance with their social hierarchy and hunting behaviour. Algorithm 2 shows the steps of BGWO.

3.2.2.1. Encircling prey. Grey wolves hunt by encircling their victim. It is written as equation (2) and (3).

$$(2) \quad B = |c \cdot x_p(t) - x(t)|$$

$$(3) \quad x(t+1) = x_p(t) - S \cdot B,$$

where t denotes the current iteration, and S represent coefficient vectors, x_p denotes the prey's position vector. The vector A is calculated as

$$S = 2a \cdot r_1 - a; \quad c = 2 \cdot r_2,$$

Algorithm 2: BGWO

Input: Population Size PS, Input features N, Superior search agent X_α , second superior search agent X_β , third superior search agent X_γ , max_iter t

Output: Optimal subset g^* , Accuracy

BEGIN

 Initialize PS, X_α , X_β , X_γ

 Initialize c

 While $c < t$

 for each feature n in population N

 Update the position of the current search feature using eq. (6)

 end for

 Compute the fitness g^*

 Update X_α , X_β and X_γ

$t = t + 1$

 end while

 Return X_α

END

where r_1, r_2 are random vectors in the range $[0, 1]$, and the components of a are linearly decreasing from the range of 2 to 0.

3.2.2.2. *Hunting.* Grey wolves are capable of tracking and encircling their prey. The Alpha wolf leads the hunt. The wolf's beta and delta siblings take after him. The x wolves come after these first three wolves.

$$(4) \quad B_\alpha = |c_1.x_\alpha - x|; \quad B_\beta = |c_1.x_\beta - x|; \quad B_\gamma = |c_1.x_\gamma - x|$$

$$(5) \quad x_1 = x_\alpha - S_1.(B_\alpha); \quad x_2 = x_\beta - S_2.(B_\beta); \quad x_3 = x_\gamma - S_3.(B_\gamma)$$

$$(6) \quad x_i(t+1) = x_1 + x_2 + x_3/3.$$

3.2.3. *Genetic optimization algorithm.* Each solution in the population is represented as a string of genes, where each gene corresponds to a particular decision variable. Algorithm 3 describes the BGA steps.

3.2.3.1 *Selection, crossover and mutation.* In selection step, a subset of the existing population of solutions is chosen to produce the upcoming generation of solutions. The fitness value of each solution is determined by how well the solution fits the optimization requirements. The chosen solutions are mixed to produce new offspring solutions during the crossover phase. This is done by exchanging segments of genes to develop fresh offspring solutions from parent solutions. Mutation is usually applied in a small number of genes within a solution, and the changes introduced are typically small. It helps to avoid premature convergence to a sub-optimal solution and can help to explore new regions of the search space.

Algorithm 3: BGA

Input: Input features N , Population Size PS , max_iteration t

Output: optimal subset g^* , Accuracy

BEGIN

 Generate initial population of n features compute the fitness value of each feature

 Initialize Counter as 0

 while (counter < t):

 select pair of features from initial population

 Apply crossover operation on selected features

 Apply mutation on the offspring

 Replace old features population with newly generated

 Increment the counter by 1

 end while

END

3.2.4. *Binarized Cuckoo optimization algorithm.* The BCOA algorithm imitates the behaviour of some cuckoo bird species that lay their eggs in the nests of other bird species, a practice known as brood parasitism. The host birds then

incubate the cuckoo eggs, which eventually hatch into cuckoo chicks that out-compete the host chicks for food. The BCOA algorithm works by maintaining a population of candidate solutions, represented by nests. Each nest corresponds to a potential solution to the optimization problem. The algorithm repeatedly creates new nests and replaces less-fit nests with new nests in an effort to find the best answer.

$$(7) \quad x_i^{t+1} = x_i^t + step_size * direction_i * magnitude,$$

where x_i^{t+1} is the new solution candidate, x_i^t is the current solution candidate, $step_size$ is a scalar value that determines the size of the step taken by the algorithm in each iteration and $direction_i$ is a vector that determines the direction of the step taken.

Algorithm 4: BCOA

Input: Input features N , Population Size PS , $step_size$ ss , magnitude m , $max_iteration$ t

Output: optimal subset g^* , Accuracy

BEGIN

 Generate initial population of n features

 Compute the fitness value of each feature

 Initialize Counter as 0

 while (counter < t):

 for each nest do

 Generate a step of magnitude $ss*m$ in the direction of $direction_i$ along the i th dimension

 Update the i th dimension of the solution vector using (7)

 end for

 end while

 return the optimal subset

END

3.2.5. Binarization and fitness function. In binarization, a threshold value is set and any continuous variable value above or equal to the threshold is assigned a binary value of 1, while any value below the threshold is assigned a binary value of 0. A fitness function is used to evaluate the quality suitability of a solution or candidate solution in an optimization or search problem.

3.3. Disease classification. In the context of feature selection, classifiers are useful for evaluating the importance of different features in predicting the class labels. Decision tree classifiers build the tree by iteratively choosing the optimal feature to divide the data into two or more subsets based on a splitting criterion like information gain or Gini measures. The KNN classifier's performance can be enhanced by feature selection, which can be used to decrease the number of features. By focusing on the most informative features and lowering the dimensionality of the input space, Gaussian classifier-based feature selection increases

Binarization $\beta(N)$	Fitness Function
Input: Solution vector s , input features x , target variable y	Begin
Output: Accuracy	$R \leftarrow$ count non zero entries in s
Initialize counter as 0	$X \leftarrow$ empty matrix with shape $\text{len}(x)$
Begin	for i in range $(0, \text{len}(s) - 1)$ do
while counter $< \text{len}(N)$	if $s[i] == 1$ then do
$r = 1 / (1 + \text{math.exp}(-10 * (x - 0.5)))$	$X[:, \text{count}] \leftarrow X[i]$
if $r < N[i]$	Count $\leftarrow$ count + 1
binarize the array	end if
$a[i] = 1$	end for
else	Split data into training and testing sets
Continue	Create model classifier
Increment the counter	Fit the classifier
by 1	Predict the model
end while	Calculate accuracy
End	Return accuracy
	End

the accuracy and effectiveness of classification models. SVM based feature selection is done by selecting a subset of features that results in the best classification performance and reducing computational complexity.

3.4. Performance evaluation. The accuracy, precision and recall of the outcome is used to estimate the model's performance.

$$Accuracy = \frac{TP_{PCOS} + TN_{PCOS}}{TP_{PCOS} + TN_{PCOS} + FP_{PCOS} + FN_{PCOS}};$$

$$Precision = \frac{TP_{PCOS}}{(TP_{PCOS} + FP_{PCOS})}; \quad Recall = \frac{TP_{PCOS}}{(TP_{PCOS} + FN_{PCOS})},$$

where True positive (TP_{PCOS}) is a solution where classifier classifies the PCOS positive as PCOS positive. True negative (TN_{PCOS}) is a solution where classifier classifies PCOS negative as PCOS negative. False positive (FP_{PCOS}) is a solution where classifier improperly classifies the PCOS positive as PCOS negative. False Negative (FN_{PCOS}) is a solution where classifier improperly classifies PCOS positive as PCOS negative.

4. Results and discussion. Four alternative optimization techniques are tested using four traditional classifiers like Decision Tree, KNN, Gaussian, SVM. The study also provides a table of common reduced features selected by the optimization algorithms. For example, the BGWO selected features like cycle, marriage status, number of abortions, weight gain, hair growth, hair loss, pimples, fast food, follicle number (R). These common features are obtained by applying optimization algorithm followed by fusion technique. Figure 2 compares the accuracy,

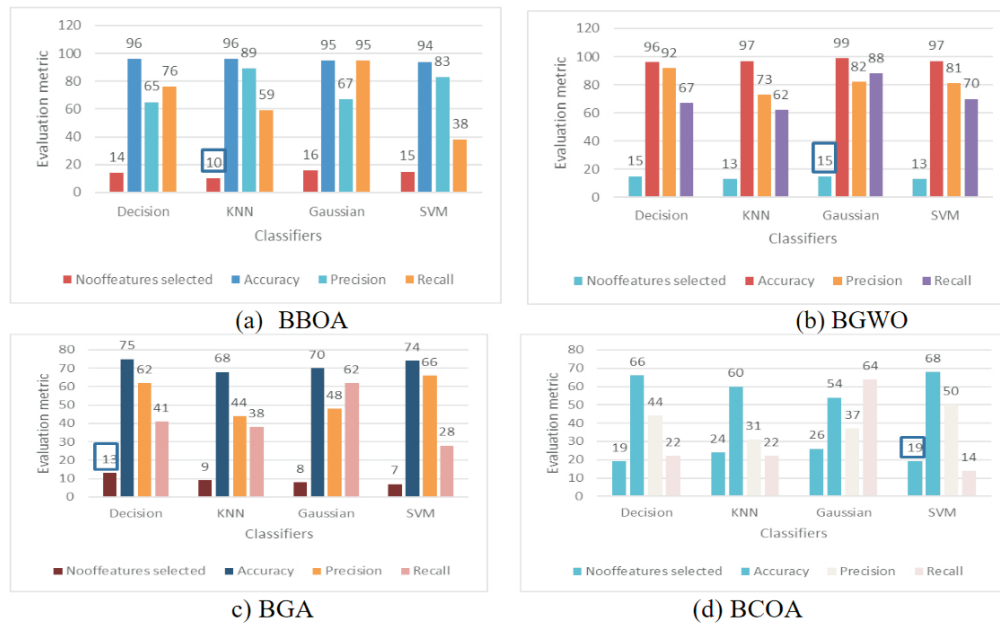


Fig. 2. Comparison of different optimizers with No. of features, accuracy, precision and recall

number of selected features, precision and recall. Figure 3 shows the confusion matrix for all optimization algorithms with four classifiers.

T a b l e 1

Common reduced features selected by the optimization algorithms

Attributes before optimization	Algorithms	Attributes after optimization
BMI, Hb, cycle, cycle length, marriage status, pregnant, No. of abortions, I betaHCG, II betaHCG, FSH/LH, waist:hip ratio, TSH AMH, PRL, PRG, RBS, weight gain, hair growth, skin darkening, hair loss, pimples, fast food, reg exercise, follicle No. (L), follicle No. (R), Avg. Fsize (L), Avg. Fsize (R), endometrium, PCOS (yes or no)	BBOA	BMI, cycle, No. of abortions, waist:hip ratio, skin darkening, pimples, fast food, follicle No. (L), follicle No. (R)
	BGWO	cycle, marriage status, No. of abortions, weight gain, hair growth, hair loss, pimples, fast food, follicle No. (R).
	BGA	BMI, Hb, cycle, pregnant, II betaHCG, TSH, AMH, weight gain, hair loss, pimples, fast food, Avg. F size (L), endometrium
	BCOA	BMI, marriage status, No. of abortions, I betaHCG, FSH/LH, AMH, PRG, pimples, fast food, follicle No. (R), Avg. F size (L), endometrium

Conclusion. The main contributions of the paper could be summarized as follows: 1. Demonstrating the significance of feature selection for optimized



Fig. 3. Confusion matrix for various optimization algorithms

PCOS diagnosis. 2. Optimized subset of features for more accessible PCOS diagnosis. 3. Interesting methodology combining several classifiers for diversified evaluation of the feature selection algorithms used. According to our analysis, the BGWO optimization algorithm outperformed the other algorithms in terms of classification accuracy, while the BGA was the most efficient in terms of providing minimal features. The actual translational benefit lies in the possibility of findings of this paper to be used by healthcare professionals to develop more accurate and efficient diagnostic tools for PCOS disease. This work can be further extended to real time hospital records such as electronic health records, scan report, health surveys, etc., and analyse the performance of the optimization algorithms.

REFERENCES

- [1] BULSARA J., P. PATEL, A. SONI, S. ACHARYA (2021) A review: Brief insight into PCOS, Endocrine and Metabolic Sci., **3**, 1–7.
- [2] RANI R., Y. A. HAJAM, R. KUMAR, R. A. BHAT et al. (2022) A landscape analysis of the potential role of polyphenols for the treatment of PCOS, Phytomedicine Plus, **2**(4), 1–21.

- [3] PREETI C. et al. (2021) Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS, International Conference on Communication information and Computing Technology (ICCICT), 7 pp, <https://ieeexplore.ieee.org/document/9510101>
- [4] DENNY A., A. RAJ, A. ASHOK, M. RAM (2019) i-Hope: Detection and Prediction System for PCOS Using Machine Learning Techniques: IEEE Region 10 Int. Conf., 673–678.
- [5] SONI P., S. VASHISHT (2019) Image Segmentation for detecting PCOS using Deep Neural Networks, Int. J. Comp. Sci. Eng., **7**(3), 534–537.
- [6] NEETHA et al. (2020) Prediction of PCOS with Clinical Dataset Using A Novel Hybrid Data Mining Classification Technique, Int. J. Adv. Res. Eng. Tech. **11**(6), 1872–1881.
- [7] DESHPANDE S. S. S., A. WAKANKAR (2014) Automated Detection of PCOS Using Follicle Recognition: IEEE Int. Conf., 1341–1346.
- [8] MADHUMITHA J., M. KALAIYARASI, S. RAM (2021) Automated PCOS Identification with Follicle Recognition, IEEE 3rd Int. Conf. on Sig. Proc. and Comm., DOI:10.1109/ICSPC51351.9451720.
- [9] DASTIDAR B. G. (2012) Automated Screening of PCOS using Machine Learning Techniques: IEEE Int. Conf., 1–5.
- [10] OTHMAN M. S., S. R. KUMARAN, L. M. YUSUF (2020) Gene Selection Using Hybrid Multi-Objective Cuckoo Search Algorithm with Evolutionary Operators for Cancer Microarray Data, IEEE Access, 8: 186348–186361.
- [11] ALAM T., S. QAMAR, A. DIXIT, M. BENAIDA (2020) Genetic Algorithm: Reviews, Implementations, and Applications, Int. J. of Eng. Pedagogy, 1–19.
- [12] TUBISHAT M., M. ALSWAITTI, S. MIRJALILI, M. AL-GARADI (2020) Dynamic Butterfly Optimization Algorithm for Feature Selection, IEEE Acc., **8**, 194303–194314.
- [13] RAJALAXMI R. R., S. MIRJALILI, E. GOTHAI, P. NATHESAN (2021) Binary Grey Wolf Optimizer with Mutation and Adaptive K-nearest Neighbour for Feature Selection in Parkinson's Disease Diagnosis, Knowl. Based Syst., 1–18.
- [14] HASSAN M. M., T. MIRZA (2020) Comparative Analysis of Machine Learning Algorithms in Diagnosis of PCOS, Int. J. of Comp. App., **175**(17), 42–53.
- [15] VEDPATHAK S., V. S. THAKRE (2020) PCO care: PCOS Detection and Prediction using Machine Learning Algorithms, Biosci. Biotech. Res. Comm., **13**(14), 240–244.
- [16] BHARATI S., P. PODDER, M. R. HOSSAIN MONDAL (2020) Diagnosis of PCOS Using Machine Learning Algorithms, IEEE Reg. 10 Sym., 1486–1489.
- [17] INAN M. S. K., R. E. ULFATH, R. HASAN (2021) Improved Sampling and Feature Selection to Support Extreme Gradient Boosting for PCOS Diagnosis, IEEE Ann. Comp. and Comm. Workshop and Conf., 1046–1050.
- [18] MEHR H. D., H. POLAT (2022) Diagnosis of PCOS through different machine learning and feature selection techniques, Health and Technology, **12**(2), 137–150.

Mepco Schlenk Engineering College
Sivakasi – 626005
Tamlinadu, India
e-mail: santhicse@mepcoeng.ac.in
nidhyan@mepcoeng.ac.in