Journal of Education for Pure Science- University of Thi-Qar
Vol.12, No2 (December, 2022)
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

DOI: http://doi.org/10.32792/utq.jceps.12.02.08

Detection of PCOS Based on Genetic Algorithm Coupled with SVM

Najlaa Nasralaa Faris and Firsas Saber Miften

University of Thi-Qar, College of Education for Pure Sciences, Iraq

Received 14/8/2022 Accepted 1/9/2022 Published 01/12/2022

Abstract:

PCOS (polycystic ovary syndrome) is a hormonal illness that affects women's of life from a young age. Human diagnosis of PCOS is a difficult undertaking for professionals; yet, rapid and accurate detection of PCOS could save the lives of millions of women throughout the world. Current research employs high-dimensional features, resulting in low estimation accuracy and a long execution time. However, in this research, we create a novel intelligence system based on a Genetic Algorithm linked with an SVM (AG-SVM) that uses fewer features to categorize PCOS. The dataset is preprocessed before a Genetic Algorithm is used to select the most powerful features for PCOS classification. As a consequence, seven features were selected from the feature set to represent PCOS data. The SVM is fed the selected characteristics set in order to categorize them into healthy and non-healthy segments. Our results showed that the suggested model (AG-SVM) achieved a 90% accuracy rate**.**

Keywords: support vector machine (SVM), PCOS, Genetic Algorithm.

1. Introduction:

Early prediction of chronic diseases has become a key factor to protect human life. Most of clinical studies have focused on investigating the relation between vulnerable women with health issues and Polycystic ovary syndrome (PCOS) [1]. PCOS, **is a hormonal disorder,** can affect women of childbearing age causing health issues such as pregnancy failure. There are many factors associated with PCOS that can increase the chance of geting PCOS including genetic, lifestyle, and diet. Clinical research showed that PCOS causes a hormonal imbalance that leads to severe issues for examples formation of follicles, cysts in the ovaries an imbalance in the level of androgen, increased hair growth, and weight gain[2]. According to the World Health Organization, PCOS risk can be increased in the case of obesity or genetics. Based on last published studies, PCOS have affected around 116 million women worldwide in 2012 [3,4].The global estimation of the PCOS are expected to be very high in the upcoming years [3]. Polycystic ovaries can lead to severe diseases in the long term, such as sleep apnea, type 2 diabetes, liver issues, coronary artery disease, and endometrial hyperplasia[4]. PCOS can also be associated with malignant tumor for example, cancerous diseases in the endometrium and breast [5].

Clinical treatments of PCOS include metformin therapy, luteinizing hormone, diet Plan, however, not all women can access these treatments especially in developing countries.

At this time there is no definitive test to predict PCOS according to Rotterdam criteria [6]. Normally, specialists inspect patient’s data visually to decide whether PCOS present or not based on the number of follicular cysts. This task is time consuming and gives low detection rate. In addition, it requires a high trained medical staff. Automatically, the diagnosis of PCOS normally carry out based on morphology on ultrasound, and androgenicity. Despite the rapid developments in detection techniques of PCOS, there have been a surge demands to develop more accurate methods.

A lot of efforts have been done by researchers to improve PCOS detection for example Munjal, Sanjay K., et al. [3] proposed a genetic algorithm to select the most representative features to detect PCOS. Several than one classifiers including extra trees, random forest, and decision tree were adopted in that study Tanwani, N.[7] suggested a method based on two machine learning KNN Logistic Regression. They extracted a set of features has and tested to show whether the extracted features reflected PCOS. In that study, an average of accuracy of 0.90, and 0.92 were obtained with 9 and 10 features respectively. Inan, Muhammad Sakib Khan, et al. [2]. Used Extreme Gradient Boostingto detect PCOS An accuracy rate of 95.83 was achieved. Satish, CR Nandipati, XinYing Chew, and Khai Wah Khaw[8]. evaluated different numbers of features to detect PCOS. Seven classification methods with five feature selection methods were used in that study. The extracted features were evlauted to find out the most powerful features An accuracy of 90.83% was attained., Malik Mubasher and Tabassim Mirza[9]. tested several machine learning algorithms in PCOS detection. They found that a Random Forest algorithm gave a good performance. Bharati, Subrato, Prajoy Podder, and M. Rubaiyat Hossain Mondal [10]detected PCOS based on a filter-based univariate feature selection method . In that study, a good accuracy and a low complexity time were obtained with 10 features. Denny, Amsey, et al.[11]. Identified PCOS using 8 features. Principal Component Analysis was used to minimize the extracted features a random forest classifier was utilized and it was obtained an accuracy of 89.02%. V. Thakre and et al.[12]. proposed a number of statistical parameters with five machine learning classifiers to detct PCOS. In that study, a random forest classifier was employed. Mehrotra, Palak, et al.[13]. proposed an approach to detect PCOS based on biomarkers. They designed an algorithm that included a vector formulation based on clinical and metabolic features. The statistically significant features were selected and sent to Bayesian classifier. another recent study made by Thomas, Neetha, and A. Kavitha. [14], a novel hybrid structure was proposed to determine the present of PCOS based on navies bayes and artificial neural network. .Cahyono, B., M. S. Mubarok, and U. N. Wisesty. [15] proposed a deep learning model to classify ultrasound images of PCOS. The data set used in that study consists of 40 non-PCOS data and 14 PCOS data.

Based on previous study results and recent technology development, we explored the feasibility of improving the accuracy of PCOS detection by using Genetic Algorithm based on SVM through improved deep learning feature selection. It has been found that there is no study that have investigated the relationships between the features number and PCOS detection. According to our knowledge, its first study which figures out the best combination of features to detect PCOS. The aim of this work is to design a new model for classifying ovaries as normal or abnormal. The most important achievements of this work are as follows:

1) investigated the best number of features to detect PCOS using Genetic Algorithm.

2) proposed a support vector machine (SVM) as a classifier to the detect PCOS.

The rest of the paper is arranged as follows: The in second describes the method used to diagnose PCOS. The in third section describes the process of finding the best features of PCOS The fourth section describes the process of classification of PCOS by applying SVM algorithm. The fifth section presents the results obtained by applying SVM to the PCOS dataset. The last section contains the conclusion, future work, and references.

2. Proposed method and dataset:

In this paper, we propose a new intelligence system from (AG-SVM) to classify PCOS diagnoses. The schematic diagram of the PCOS detection is shown in Figure 1. The first block is the PCOS database entry, the second block is preprocessing, and the third block is the feature selection using Genetic Algorithm to be used for classification by the SVM classifier. The samples are divided into two groups the training and the testing.

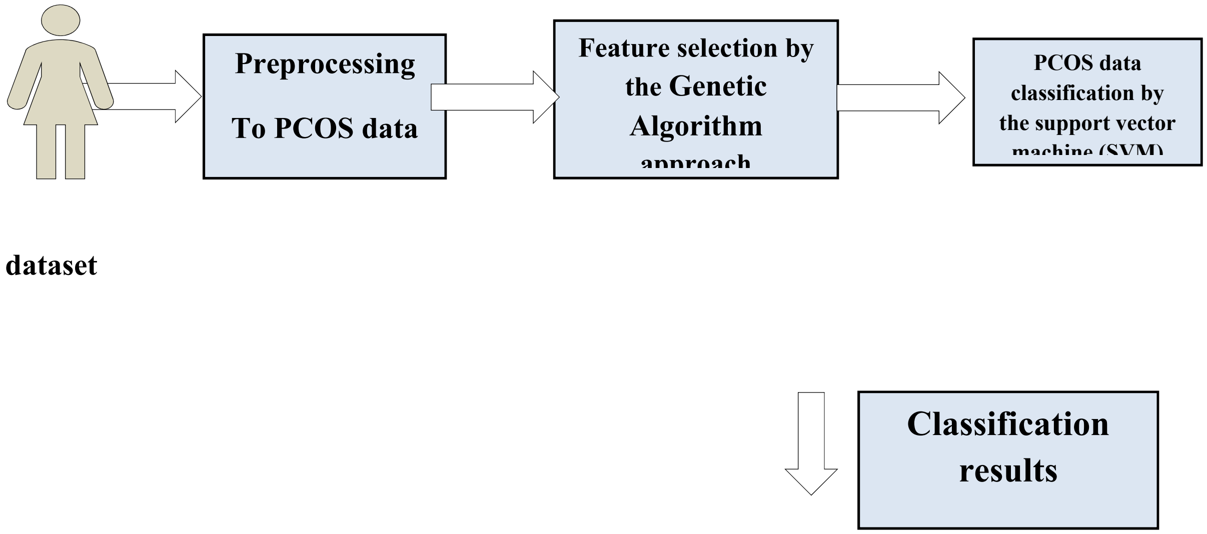

Preprocessing
To PCOS data
Feature selection by the Genetic Algorithm approach
PCOS data classification by the support vector machine (SVM)
dataset
Classification results

Fig. 1. Block diagram of the proposed method

2.1. Dataset

The dataset [16] was used in this paper to evaluate the proposed method. It has been collected by Prasoon Kottarathil (data scientist in Muriyad, Kerala, India) from 10 hospitals across Kerala, India in 2020. A total 541 subjects were involved in that study. only 177 women were diagnosed with PCOS. Their demography information was recorded as: aged 20-48, height 134-171, and weight between 31-108. Includes 42 features, The dataset included 41 attributes for each subject. We found that only 14 features were effective in the PCOS detection. More details regarding feature selection were presented in the experimental results section. Table 2 reports the selected features of PCOS.

Table 1: List of input attributes used in diagnosis of PCOS

Rank	Feature	Rank	Feature	Rank	Feature
1	PCOS (class label)	15	II beta-HCG (mIU/mL)	29	hair growth(Y/N)
2	Weight (Kg)	16	FSH (mIU/mL)	30	Skin darkening (Y/N)
3	Height (Cm)	17	LH (mIU/mL)	31	Hair loss(Y/N)
4	BMI	18	FSH/LH	32	Pimples(Y/N)
5	Blood Group	19	Hip(inch)	33	Fast food (Y/N)
6	Pulse rate(bpm)	20	Waist(inch)	34	Reg Exercise (Y/N)
7	RR (breaths/min)	21	Waist: Hip Ratio	35	BP _Systolic (mmHg)
8	Hb(g/dl)	22	TSH (N/AbN)	36	BP _Diastolic (mmHg)
9	Cycle(R/I)	23	AMH (ng/mL)	37	Follicle No. (L)
10	Cycle length(days)	24	PRL (ng/mL)	38	Follicle No. (R)
11	Marriage Status (Yrs.)	25	Vit D3 Deficiency(Y/N)	39	Avg. F size (L) (mm)
12	Pregnant(Y/N)	26	PRG (ng/mL)	40	Avg. F size (R) (mm)
13	No. of abortions	27	RBS (mg/dl)	41	Endometrium (mm)
14	I beta HCG (mIU/mL)	28	Weight gain(Y/N)	42	Age (yrs.)

Body Mass Index (BMI), The level of follicle stimulating hormone (FSH), Luteinizing hormone (LH), Blood Pressure (BP), Hemoglobin (Hb), beta -Human Chorionic Gonadotropin (HCG), Thyroid Stimulating Hormone (TSH), Prolactin (PRL), Progesterone (PRG), Anti-Mullerian Hormone (AMH), Random Blood Sugar (RBS)

2.2 Data preprocessing:

In this section, previously collected data has been processed. We removed unwanted columns from the dataset and removed redundant features. Additionally, the data set has been standardized into numerical forms. it is converted into a matrix of 538X42 where 538 refers to samples and 42 indicate the features used for each sample to diagnose PCOS using the proposed model (AG-SVM).

2.3 Feature selection:

Genetic algorithms (GAs) are best known for their ability to efficiently search large spaces about which little is known a priori. Since genetic algorithms are relatively insensitive to noise, they seem to be an excellent choice for the basis of a more robust feature selection strategy for improving the performance of PCOS classification system. In this section we describe this approach in more detail.

Genetic algorithms:

Genetic algorithms (GAs), a type of inductive learning strategy, are adaptive search techniques that outperform a variety of random and local search methods [19]. This is accomplished through their ability to use accumulating information about a previously unknown search space to bias subsequent searches into promising subspaces. GAs are ideal for applications where domain knowledge and theory are difficult or impossible to provide because they are essentially a domain independent search technique [19]. The main challenges in applying GAs to any problem are choosing an appropriate representation and an appropriate evaluation function. See [20] for a detailed description of both of these issues in the context of the feature selection problem. The main goal of the feature selection problem is to represent the space of all possible subsets of the given feature set. The simplest form of representation is binary representation, in which each feature in the candidate feature set is treated as a binary gene, and each individual is composed of a fixed-length binary string representing some subset of the given feature set. A l-dimensional binary feature vector X corresponds to an individual of length l, where each bit represents the elimination or inclusion of the associated feature. Then, $X_i = 0$ denotes

2.4 Support Vector Machines (SVM):

which is a method for classifying both linear and nonlinear data. A support vector machine (SVM) [17] is one of the most usable binary classification models.

In a nutshell, an SVM is a machine learning algorithm that operates like this [18]:

1. The original training data is transformed into a higher dimension using a nonlinear mapping method.
2. It looks for the linear optimum separation hyperplane (a "decision border" dividing tuples of one class from those of another) within this new dimension.
3. Data from two classes may always be separated by a hyperplane if a suitable nonlinear mapping to a sufficiently high dimension is used.
4. Support vectors ("essential" training tuples) and margins are used by the SVM to locate this hyperplane (defined by the support vectors).

Suppose a set of training data belonging to two groups, the SVM uses a hyperplane to separates those two samples into two classes. The SVM is trained with a paired (feature and class) and y_i . Let a training set comprising of N classes with f attributes, the set can be represented by $\{(x_i,y_i) \mid x_i \in R^f , y_i \in \{-1, 1\}\}$

where i=(1,…,N) If the input data is linearly distinguishable, the optimization phase can be defined as follows [17]:

$$\min_{w,b} ||w||^2 \qquad (1)$$

Where:

yi(w.xi+b)≥1

w: is a weight vector.

b: is the bias term.

the hyperplane is defined by *w. x*+*b*=0.

Suppose entity in the testing set is classified using the decision function *f(x)* = *w. x*+*b*. If *f*(x) < 0. As a result, the instance falls on one side of the hyperplane. However, the entity *x* is identified and put in the first class (*y*=−1), otherwise is classified into the second class (*y*=1).

In the case of the input data is not linearly separable. The formulation is updated to be a soft-margin. To tackle the misclassifications, the following optimization formula is employed

$$\min_{w,b} ||w||^2 + c \ \sum_{i=1}^{\mathrm{N}} \xi_i^2 \qquad (2)$$

Where:

yi(w . xi)≥1 –*ξi, ξi*≥ 0
w: is a weight vector.
b: is the bias term.
ξi: is a slack variable.

c: amount of error.

The soft-margin formulation is utilized to figure out the hyperplane solution to separate between maximizing the margin and the amount of error *C*.

The above formula is used with the linear classification; however, it can be applied to the nonlinear classification by using the kernel trick. The instance is mapped into some higher dimensional space based on the following formula $x \mapsto \Phi(x)$ where is $\Phi(x)$ a nonlinear mapping function.

When the dot product between two entities is required in the SVM, the kernel trick is employed which is defined based on the following formula

$$\mathrm{K}\ (x_i, x_j) = \Phi(x_i).\ \Phi(x_j) \qquad (3)$$

Where Φ(*xi*) and Φ(*xj*) is a nonlinear mapping function. The SVM optimization function is computed based on the following formula:

$$\max_{\alpha} \sum_{i=1}^{\mathrm{N}} \alpha i - \frac{1}{2} \sum_{i,j} \alpha i \alpha j y i y i \mathrm{K}(xi, xj) \qquad (4)$$

Where:

$$\sum_{i=1}^{\mathrm{N}} \alpha i y i = 0, \alpha i \ \geq 0$$

xi, xj : are the th and th input features.
yi, yj: are the class labels *of xi* and *xj***.**
K: kernel trick function.

3. Performance evaluation metrics:

The performance of the proposed model is tested using several metrics such as accuracy and recall value, confusion matrix, and F score.

3.1Confusion matrix:

The confusion matrix is a used to assess the proposed model's performance. The matrix compares the least square support vector machine model's predictions to the actual target values. It gives us a full picture of our classification model's performance as well as the kinds of errors it makes. The matrix contains 'True Negative (TN)' ,and 'False Negative (FN)', and 'True Positive (TP)', and 'False Positive values (FP)'.
Recall is the number of true positives divided by the number of actual positives

Recall $= \frac{TP}{TP+FN}$. (5)
Precision is a measure of convergence of compatible readings.
Precision is the number of true positive outcomes divided by the number of expected positive outcomes. That is, the percentage of all positive ratings that we get correctly.

Precision $= \frac{TP}{TP+FP}$. (6)

F-measure is the harmonic average of accuracy and recall

F-measure$= 2 \times \frac{\textbf{Precision}*\textbf{Recall}}{\textbf{Precision}+\textbf{Recall}}$ (7)

Specificity is the ability of the test to identity individuals without disease.

Specificity$= \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}}$ **(8)**

Accuracy is the measurement of the true value from the measured value.

$$\textbf{accuracy} = \frac{\text{Number of successfully classified objects}}{\text{total number of objects in the test set}} \quad . \qquad (9)$$

4. Experimental Results and Discussion:

Seven features were retrieved and sent into the SVM in this experiment. All of the retrieved traits were put to the test individually, and the results were kept track of. The detection rate for each feature is reported in Table 1. When compared to other features, the Follicle feature has the highest detection rate. Weight growth, on the other hand, had the least detection rate.

Our study included a total of 538 participants, 364 of whom were in the control group, and 177 of whom were diagnosed with PCOS. By deleting unnecessary columns and standardizing the data, we were able to examine the dataset. The Genetic Algorithm was used to select the most powerful traits. To optimize the SVM, the RBF kernel function was chosen. The SVM has two critical parameters, and 2, which should be chosen carefully to achieve the desired results. The SVM was trained with several combinations of the parameters and 2 to get the best results. The samples were separated into two groups: 438 for training and 100 for testing.

TABLE 2. RANKING OF BEST 6 FEATURES ARRANGED IN DESCENDING ORDER

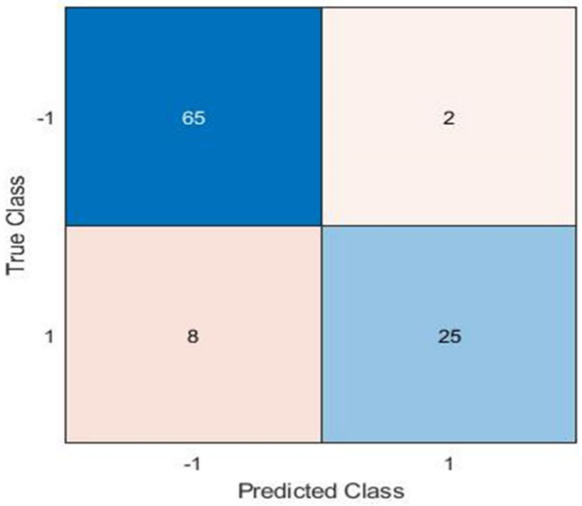
Features
RR
Cycle length(days)
hair growth
Pimples
Fast Food
Follicle No. (R)

In this paper, all seven features were used to detect PCOS, and the features were fed into SVM. SVM performance was evaluated using accuracy, recall, presession, and f-score. Table 3 displays the proposed model's detection rate.

Table 3 – Performance the proposed **AG-SVM** model

S.N.	Classifier	Accuracy	Recall	Prec.	Specificity	F-Measure
1	SVM	0.90	0.757	0.92	0.97	0.83

The proposed model's confusion matrix was computed. Out of 100 samples, 90 (65+25) were correctly classified, while 10 (2+8) were incorrectly classified.



When 21 features were used (Follicle No. (R), Follicle No. (L), BP _Diastolic, Reg. Exercise, Pimples, Hair loss, Skin darkening, hair growth, Weight gain, PRG, VitD3 Deficiency, Waist: Hip Ratio, Waist, FSH/LH, LH, FSH, Cycle length(days), Cycle, Pulse rate, BMI, Weight), the detection rate dropped to 80The detection rate was enhanced by 12% with 7 and 10 features, and accuracy was measured at 90 percent and 88 percent for 7 and 10 features, respectively. Different SVM kernels were tested in this study. The accuracies obtained using kernel functions are listed in Table 4. When compared to other kernel functions, the RBF kernel has the highest accuracy of 90 percent. Poiy-kernel, on the other hand, had the lowest accuracy rate at 80 percent.

Table 4: Performing the proposed method with different kernel types

Type kernel	Accuracy	Sensitivity	Precision	Specificity	F-measure
Lin_ kernel	0.88	0.72	0.92	0.95	0.81
Poly-kernel	0.80	0.70	0.90	0.90	0.80
RBF-kernel	0.90	0.757	0.92	0.97	0.83

Note: Lin = Linear kernel; Poly=Polynomial kernel; RBF=Radial Basis Function kernel.

Comparing previous studies with our method:

The performance of proposed GA-SVM in classifying PCOS patients is comparable with the results reported in the literature. In this paper for LS-SVM algorithms, only seven features are used which is less than that of the previous research papers. Therefore, the classification of the proposed ga-SVM used in this work require less computation time compared to the existing research works. Due to the common metabolic disorders and long-term complications in PCOS patients, the number of studies of metabolic profiles characterizing different phenotypes, revealing pathways and identifying biochemical markers of PCOS are increasing, Blood sample testing has been widely applied in the clinical diagnosis. In Table 6

Table 6: Comparing previous studies with our method

Authors	Classifier	Number of Features	Recall	Prec.	Accuracy
Proposed method	Support Vector Machine Algorithm	7	0.757	0.92	0.90
Namrata Tanwani [7]	logistic regression	10	0.82	0.93	0.92
Bharati et al. [10]	hybrid random forest and logistic regression	10	0.90	0.89	0.91
V. Thakre and et al. [12]	random forest	30	0.80	0.891	0.909
Denny, Amsey, et al. [11]	random forest	12	0.741	0.958	0.89

5. Conclusion:

In this paper, we developed a new intelligence system to detect PCOS based on genetic algorithm coupled with SVM (GA-SVM). A total of 541 samples which were collected from ten different Indian hospitals, were used in this paper. The low number of features and the short inference time showed the feasibility of the proposed model for real-time processing on clinical systems. The results showed that curacy rate was increased when 7 features were adopted. The SVM gained a high detection rate compared with other classification models. We will validate the proposed model with aa huge dataset.

References:

1. Krishnaveni, V., *A Roadmap to a Clinical Prediction Model with Computational Intelligence for PCOS.* International Journal of Management, Technology and Engineering, 2019. 9(2): p. 177-185.
2. Inan, M.S.K., et al. *Improved Sampling and Feature Selection to Support Extreme Gradient Boosting For PCOS Diagnosis*. in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. 2021. IEEE.

3. Munjal, A., R. Khandia, and B. Gautam, *A machine learning approach for selection of Polycystic Ovarian Syndrome (PCOS) attributes and comparing different classifier performance with the help of WEKA and PyCaret.* Int. J. Sci. Res, 2020: p. 1-5.
4. Sumathi, M., et al. *Study and detection of PCOS related diseases using CNN*. in *IOP Conference Series: Materials Science and Engineering*. 2021. IOP Publishing.
5. Soni, P. and S. Vashisht. *Exploration on polycystic ovarian syndrome and data mining techniques*. in *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*. 2018. IEEE.
6. Zhang, X., et al., *Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening.* Molecular and Cellular Endocrinology, 2021. 523: p. 111139.
7. Tanwani, N., *Detecting PCOS using Machine Learning.* IJMTES | International Journal of Modern Trends in Engineering and Science 2020. 07 (01 2020).
8. Satish, C.N., X. Chew, and K.W. Khaw, *Polycystic Ovarian Syndrome (PCOS) classification and feature selection by machine learning techniques.* 2020.
9. Hassan, M.M. and T. Mirza, *Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome.* International Journal of Computer Applications. 975: p. 8887.
10. Bharati, S., P. Podder, and M.R.H. Mondal. *Diagnosis of polycystic ovary syndrome using machine learning algorithms*. in *2020 IEEE Region 10 Symposium (TENSYMP)*. 2020. IEEE.
11. Denny, A., et al. *I-HOPE: detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques*. in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. 2019. IEEE.
12. Vaidehi Thakre1*, S.V., K.T. , and a.S. Sonawani4, *PCOcare: PCOS Detection and Prediction using Machine*

Learning Algorithms. Biosc.Biotech.Res.Comm. Special, 2020. 13(14).

13. Mehrotra, P., et al. *Automated screening of polycystic ovary syndrome using machine learning techniques*. in *2011 Annual IEEE India Conference*. 2011. IEEE.
14. Thomas, N. and A. Kavitha, *PREDICTION OF POLYCYSTIC OVARIAN SYNDROME WITH CLINICAL DATASET USING A NOVEL HYBRID DATA MINING CLASSIFICATION TECHNIQUE.* Technology, 2020. 11(11): p. 1872-1881.
15. Cahyono, B., M. Mubarok, and U. Wisesty. *An implementation of convolutional neural network on PCO classification based on ultrasound image*. in *2017 5th International Conference on Information and Communication Technology (ICoIC7)*. 2017. IEEE.
16. https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos/version/1. 2020.
17. C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

18. J. Han, M. Kamber, and J. Pei, “Introduction,” in *Data Mining*, 2012, pp. 1–38.19. Esen, H., et al., *Modelling of a new solar air heater through least-squares support vector machines.* Expert Systems with Applications, 2009. 36(7): p. 10673-10682.
19. De Jong, K. “Learning with Genetic Algorithms : An overview,” Machine Learning Vol. 3, Kluwer Academic publishers, 1988.

20. Vafaie, H., and De Jong, K.A., "Improving the performance of a Rule Induction System Using Genetic Algorithms," Proceedings of the First International Workshop on MULTISTRATEGY LEARNING, Harpers Ferry, W. Virginia, USA, 1991.