Journal of Education for Pure Science- University ofThi-Qar
Vol.13, No.1 (March., 2023)
Website:jceps.utq.edu.iq *Email:jceps@eps.utq.edu.iq*

DOI: http://doi.org/10.32792/utq.jceps.10.01.01

Proposed Model for Detection of PCOS Using Machine learning methods and Feature Selection

Najlaa Nsrulaah Faris and Firsas Saber Miften

Firas@utq.edu.iq; najlaansrulaah@gmail.com

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

Received 3/7/2022, Accepted 19/5/2022, Published /March/2023

Abstract:

PCOS (Polycystic Ovary Syndrome) is an endocrine disorder that affects women of reproductive age. Once diagnosed, the condition cannot be cured, but treatment can help relieve its symptoms. Although the exact cause of PCOS is unknown, certain risk factors do exist. Obesity, insulin resistance, blood pressure, depression, and inflammation are all factors that contribute to this syndrome. While the symptoms include hirsutism, oligo-ovulation, acne, heavy bleeding, and skin darkening. Using the causes and symptoms, a model is created that accepts them as features and outputs the presence or absence of this condition. Machine learning models used for classification include Linear SVM, Gaussian SVM, Decision Tree, Nave Bayes, KNN1, KNN3, and KNN5. The goal of building multiple models is to find the best one for a given dataset within the known scope of knowledge.

Keywords: PCOS, Genetic algorithm, Machine learning, KNN, SVM

Introduction:

Polycystic ovary syndrome or PCOS, is a widely known endocrine disorder in women distinguished by hyperandrogenism during reproductive age [1]. This syndrome was initially reported by Stein and Leventhal in 1935 and it was known for a long time as the Stein-Leventhal syndrome [2]. Women with PCOS often could suffer from menstrual problems and infertility issues [3]. Moreover, it may contribute to long-term health problems like diabetes and heart disease, cancer of the uterus and mood disorders [4]. The exact causes of PCOS are very complex and still not clear. There are many suggested etiologies for PCOS, but it is not fully supported, in general, this hormonal imbalance consists of a combination of excess androgen and insulin resistance[5]. In addition to environmental and genetic factors that contribute to this hormonal imbalance, all of these causes the development of PCOS[6]. Symptoms and signs vary among women with PCOS, it includes a combination of hyperandrogenism (hirsutism, alopecia, acne, high blood testosterone), obesity and severe menstrual irregularity [7]. According to the "Rotterdam European Society for Human Reproduction and Embryology and the American Society for Reproductive Medicine (ESHRE/ASRM) Sponsored PCOS Consensus Workshop" (Rotterdam ESHRE/ASRM, 2004), the presence of two of the three diagnosis criteria: "oligomenorrhoea, polycystic ovaries morphology, and hyperandrogenemia" [8].

The various machine learning classification methods have been used for the detection of various diseases such as breast cancer, heart and ovarian, etc. [13]. Some of the classification algorithms used for the prediction of PCOS dataset as **Denny, Amsey, et al (2019).** [9], 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. Thakre1 and et al**. **(2020)** [10] , proposed a number of statistical parameters with five machine learning classifiers to detct PCOS. In that study, theyfound that the random forest classifier is more reliable than the others, with an accuracy of 90.9%. **Munjal, Sanjay K., et al**. **(2020)** [12] , proposed a genetic algorithm to select the most representative features to detect PCOS. Several than one classifier including extra trees, random forest, and decision tree were adopted in that It was the best result Extra trees with 88% accuracy. **Silva, I. S., et al**. **(2021)** [11], they used 72 patients with PCOS and 73 healthy women. Focusing on only 10 features. For data classification, they suggested BorutaShap method, followed by the Random Forest algorithm, which gave an accuracy of 86%.

2. DIFFERENT MACHINE LEARNING ALGORITHMS

There are many algorithms used in Machine Learning for the prediction of target values based on various input values. The algorithms taken under use for the prediction of PCOS are as following: -

2.1 Support Vector Machine (SVM)

A classification technique, where each data item is plotted in n- dimensional space as a point. It comes under supervised machine learning model. The support vectors lie near the margin of the classifier.

2.2 Naïve Bays

One of the powerful classification methods evolved on Bays' theorem with independence between predictors as an assumption. The Naïve bays classifier assumes that there is no relation between the presences of a particular feature over the other features. The status of a particular feature does not affect the status of another feature.

2.3 Decision tree

Decision tree is a tree structure which looks similar as flow chart. Every node in the tree represents the test of an attribute, every branch represents the output of the test, and every leaf means a class or distribution of classes. The top node is root node, from root node to one leaf consists a classification rule. So, decision tree is easy to be transferred to classification rules. It has many algorithms, but the main idea is using from top to bottom induction method. And the most important part is choosing which attribute to be the node as well as the evaluation of whether the tree is correct.

2.4 KNN

Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all the training tuples are stored in an n-dimensional pattern space. When given an unknown tuple, K-NN classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are K-NN of the unknown tuple.

3. OBJECTIVES

A. To diagnose PCOS based on clinical symptoms associated with the using popular machine learning algorithms on random data samples.

B. To compare performance of different algorithms and determine the best possible algorithm among them.

4. METHODOLOGY OF THE RESEARCH

The sound methodology is the key of a successful research. The methodology adopted to carry on this study is shown in the figure1.

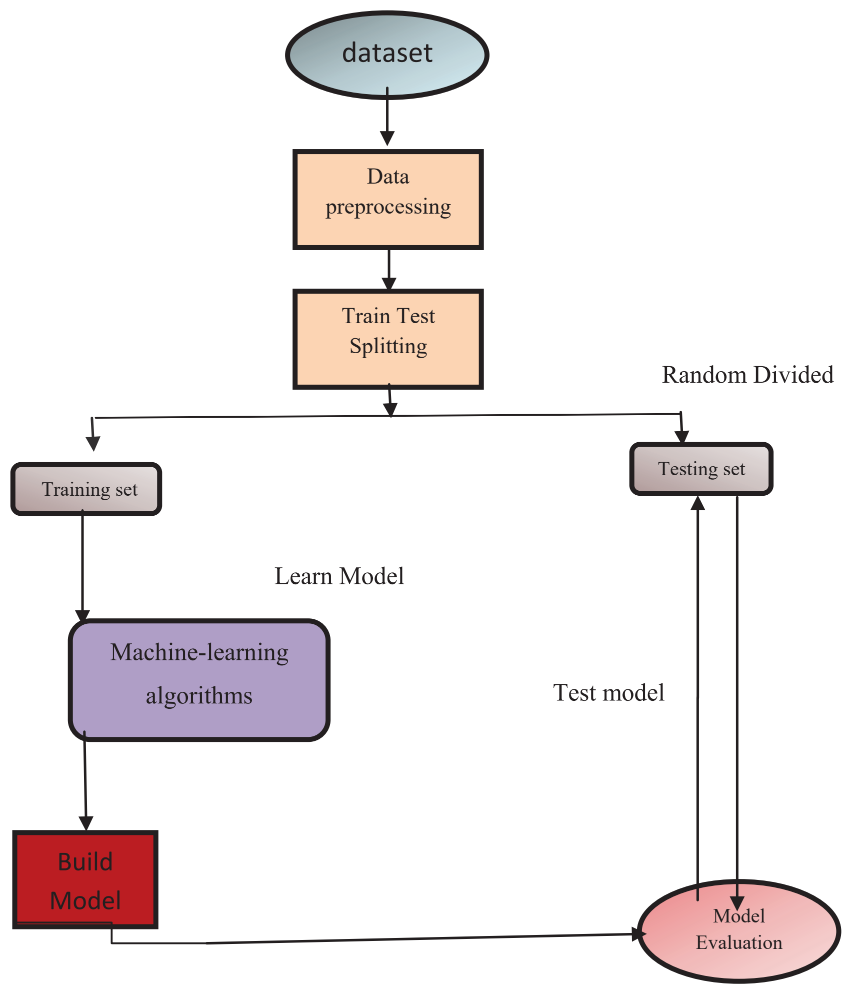


Figure (1): models (Machine learning algorithms) implemented

4.1 Data Sources and Attributes Description

The PCOS dataset for this study was obtained from Kaggle [14]. From previous studies, it is clear that researchers and clinicians are using different disease datasets to study machine learning classification methods, similar to the PCOS dataset [13]. The original PCOS dataset contains 541 instances with 42 attributes, one of which is the patient file number (not taken into consideration for data analysis). Finally, there are 41 attributes in total, with 40 as input attributes and PCOS as a class label [Positive (Yes) and Negative (No)]. The dataset has an uneven distribution of class labels (i.e., 364 instances of class label = 0 and 177 instances of class label =1) and missing values. The data was collected from ten different hospitals in Kerala, India, and it now available on the Kaggle website. Continuous, nominal, and ordinal expressions The description of the attributes is shown in Table 1.

Table 1 Attributes description and their units

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

4.2 Preprocessing of data

The data is preprocessed to identify missing values before being used to diagnose PCOS via various machine learning algorithms.

4.3 Genetic algorithm-based feature selection

In this section the application of a genetic algorithm for feature selection is described Because genetic algorithms are relatively insensitive to noise. Binary genetic algorithm (GA) was used to reduce dimensionality to improve the performance of the respective classifiers the PCOS classification. The carefully selected fitness function enabled GA to reduce the classification error from KNN. the best fitness and mean fitness should be close in value as the GA reaches the termination condition. The stall generation is number of generations produced by the GA since the last upgrade of the fitness value. In this work only 7 chromosomes were extracted out of 41feature. Table 2 shows the extracted features (RR, Cycle length (days), VD3, hair growth, Pimples, Fast Food, Follicle No. (R)).

TABLE (2) RANKING OF BEST 7 FEATURES ARRANGED IN DESCENDING ORDER
Features
RR
Cycle length(days)
hair growth
Pimples
Fast Food
Follicle No. (R)
VD3

4.4 Classification

We use MATLAB to implement classification algorithms, diagnose problems, and validate model performance. To diagnose PCOS, seven machine learning algorithms were trained and evaluated on random samples of data with 42 independent variables as symptoms. The dependent variable PCOS has a strong correlation with these independent variables and has two possible values: '1' or '0.' Among the algorithms used are knn1, knn3, knn5, SVM (Support Vector Machine), Decision tree, and Nave Bayes, Linear SVM. Table 1 depicts the list of variables used in the diagnosis of PCOS.

4.5 Validation of models

Three widely used parameters, accuracy, sensitivity, and specificity, presented in Eqs. (1) to (5), to evaluate machine learning algorithms, are used in this paper to demonstrate the capability of the proposed method to diagnose PCOS.

$$\text{Recall} = \frac{TP}{TP+FN} \quad . \qquad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad . \qquad (2)$$

$$\text{F-measure} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (4)$$

$$\text{accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}} \quad . \qquad (5)$$

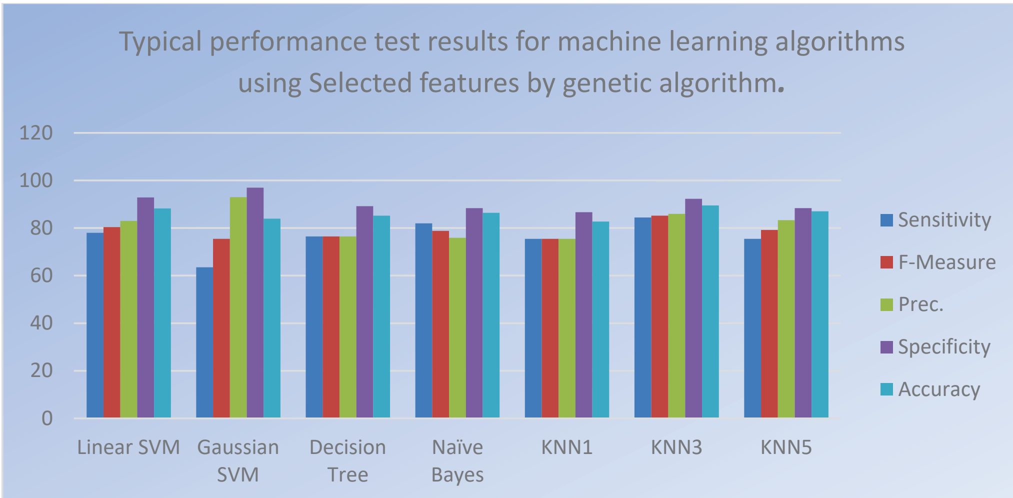
TP shows a person does not have PCOS and identified as a nonPCOS patient, and TN shows a PCOS patient correctly identified as a PCOS patient. FN shows the patient has PCOS but is predicted as a healthy person. Moreover, FP shows the patient is a healthy person but predicted as a PCOS patient.

5. RESULTS

Classifiers are fed the best features chosen by the genetic algorithm. Table 3 shows how accurate classifiers are when they are trained on the same problem. When the seven features chosen by the genetic algorithm were entered, the KNN3 algorithm and Linear SVM algorithm achieved the highest accuracy in this classifier, reaching 89.51 percent and 88.72 percent, respectively, with the specificity increasing in both. The KNN3 algorithm significantly reduces sensitivity, F-Measure, and Prec. by 84.48 percent, 85.22 percent, and 85.96 percent, respectively. Sensitivity, F-Measure, and Prec. in the Linear SVM algorithm are 78.00 percent, 80.41 percent, and 82.98 percent, respectively.

Table 3. Typical performance test results for machine learning algorithms using Selected features by genetic algorithm.

S.N.	*Network AG*	*Sensitivity (%)*	*F-Measure (%)*	*Prec. (%)*	*Specificity (%)*	*Accuracy (%)*	*Confusion Matrix* *Predict Class*			
							0	1		
1	Linear SVM	78.00	80.41	82.98	92.86	88.27	99	12	0	*True Class*
							12	39	1	
2	Gaussian SVM	63.49	75.47	93.03	96.97	83.95	104	8	0	
							11	39	1	
3	Decision Tree	76.47	76.47	76.47	89.19	85.19	96	3	0	
							23	40	1	
4	Naïve Bayes	82.00	78.85	75.93	88.39	86.42	91	14	0	
							14	43	1	
5	KNN1	75.44	75.44	75.44	86.67	82.72	98	8	0	
							9	99	1	
6	KNN3	84.48	85.22	85.96	92.31	89.51	101	8	0	
							13	40	1	
7	KNN5	75.47	79.21	83.33	88.39	87.04	99	13	0	
							9	4	1	



6. CONCLUSIONS

An early diagnosis of the disorder can aid in its treatment and management. The performance validation metrics recall, accuracy, precision, and F- statistics showed that the KNN3 algorithm performed best in diagnosing PCOS with an accuracy of 89.51 percent, followed by the Linear SVM algorithm with an accuracy of 88.27 percent. The study's future scope may include the use of various or large data sets for disease diagnosis.

REFERENCES

[1] X. Zhang *et al.*, "Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening," *Mol. Cell. Endocrinol.*, vol. 523, no. December 2020, p. 111139, 2021, doi: 10.1016/j.mce.2020.111139.

[2] K. Maheswari, T. Baranidharan, S. Karthik, and T. Sumathi, "Modelling of F3I based feature selection approach for PCOS classification and prediction," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 1, pp. 1349–1362, 2021, doi: 10.1007/s12652-020-02199-1.

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

[4] J. Madhumitha, M. Kalaiyarasi, and S. S. Ram, "Automated polycystic ovarian syndrome identification with follicle recognition," *2021 3rd Int. Conf. Signal Process. Commun. ICPSC 2021*, no. May, pp. 98–102, 2021, doi: 10.1109/ICSPC51351.2021.9451720.

[5] S. Bharati, P. Podder, and M. R. Hossain Mondal, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms," *2020 IEEE Reg. 10 Symp. TENSYMP 2020*, no. June, pp.

1486–1489, 2020, doi: 10.1109/TENSYMP50017.2020.9230932.

[6] P. Soni and S. Vashisht, “Exploration on Polycystic Ovarian Syndrome and Data Mining Techniques,” *Proc. 3rd Int. Conf. Commun. Electron. Syst. ICCES 2018*, no. Icces, pp. 816–820, 2018, doi: 10.1109/CESYS.2018.8724087.

[7] M. S. Khan Inan, R. E. Ulfath, F. I. Alam, F. K. Bappee, and R. Hasan, “Improved Sampling and Feature Selection to Support Extreme Gradient Boosting for PCOS Diagnosis,” *2021 IEEE 11th Annu. Comput. Commun. Work. Conf. CCWC 2021*, pp. 1046–1050, 2021, doi: 10.1109/CCWC51732.2021.9375994.

[8] U. N. Wisesty and J. Nasri, “Modified backpropagation algorithm for polycystic ovary syndrome detection based on ultrasound images,” in *International Conference on Soft Computing and Data Mining*, 2016, pp. 141–151.

[9] R. M. Dewi, Adiwijaya, U. N. Wisesty, and Jondri, “Classification of polycystic ovary based on ultrasound images using competitive neural network,” *J. Phys. Conf. Ser.*, vol. 971, no. 1, 2018, doi: 10.1088/1742-6596/971/1/012005.

[10] N. N. Xie, F. F. Wang, J. Zhou, C. Liu, and F. Qu, “Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network,” *Biomed Res. Int.*, vol. 2020, 2020, doi: 10.1155/2020/2613091.

[11] V. Fruh, J. J. Cheng, A. Aschengrau, S. Mahalingaiah, and K. J. Lane, “Fine particulate matter and polycystic ovarian morphology,” *Environ. Heal.*, vol. 21, no. 1, pp. 1–8, 2022, doi: 10.1186/s12940-022-00835-1.

[12] C. Neto, M. Silva, M. Fernandes, D. Ferreira, and J. Machado, “Prediction models for Polycystic Ovary Syndrome using data mining,” in *International Conference on Advances in Digital Science*, 2021.

[13] Wang, J., Jain, S., Chen, D., Song, W., Hu, C.T., & Su, Y.H., Development and Evaluation of Novel Statistical Methods in Urine Biomarker-Based Hepatocellular Carcinoma Screening. Sci Rep 8, 1 (2018) 3799.

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

[15] Azeez, R. A., Miften, F. S., & Hayawi, M. J. (2020). Epileptic EEG Signals Classification Based on Determinant of Matrix as a Feature. 4. *Journal of Global Scientific Research (ISSN: 2523-9376)*, *4*, 461-465.

[16] Al-Salman, W., Li, Y., Wen, P., Miften, F. S., Oudah, A. Y., & Al Ghayab, H. R. (2022). Extracting epileptic features in EEGs using a dual-tree complex wavelet transform coupled with a classification algorithm. *Brain Research*, *1779*, 147777.

[17] Miften, F. S. (2017). Using Quantum Particle Swarm Optimization to Enhance K-Means Clustering. *Journal of Education for Pure Science*, *7*(4).

[18] Miften, F. S., Taher, H. B., & Ajeel, K. (2017). Fractal Image Compression using Quantum PSO. *Journal of Education for Pure Science*, *7*(2).

[19] Hadi, I., & Miften, F. S. (2017). A Graph Clustering Algorithm Based on Adaptive Neighbors Connectivity. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *9*(2-11), 19-22.

[20] Miften, F. S. (2014). Upgrade PIFS to Patterns Recognition.