5th International Conference on Engineering Technology and its Applications 2022- (5thIICETA2022)

Detection of Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms

1st Dana Hdaib
Department of Biomedical Engineering
Jordan University Of Science And Technology
Irbid, Jordan
dwhdaib17@eng.just.edu.jo

2nd Noor Almajali
Department of Biomedical Engineering
Jordan University Of Science And Technology
Irbid, Jordan
nzalmajali17@eng.just.edu.jo

3rd Hiam Alquran
Department of Biomedical Engineering
Jordan University of Science and Technology
Biomedical Systems and Medical Informatics Engineering
Yarmouk University
21163 Irbid, Jordan
heyam.q@yu.edu.

4th Wan Azani Mustafa
Faculty of Electrical Engineering Technology
Univerisiti Malaysia Perlis (UniMAP)
Pauh Putra Campus
Advanced Computing Centre of Excellence
Universiti Malaysia Perlis (UniMAP)
Pauh Putra Campus
02600 Arau, Perlis, Malaysia
wanazani@unimap.edu.my

5th Waleed Al-Azzawi
Medical technical college
Al-Farahidi University
Baghdad,Iraq
Aqeel@alfarahidiuc.edu.iq

6th Ahmed Alkhayyat
Faculty of Engineering
The Islamic University
Najaf, Iraq
dr.ahmedalkhayyat85@gmail.com

2022 5th International Conference on Engineering Technology and its Applications (IICETA) | 978-1-6654-7215-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IICETA54559.2022.9888677

***Abstract*— One of the most common diseases in women of reproductive age is Polycystic Ovary Syndrome (PCOS). PCOS diagnosis can be tricky, because not everyone with PCOS has polycystic ovaries (PCO), nor does everyone with ovarian cysts have PCOS, hence the pelvic ultrasound as a stand-alone diagnosis is not sufficient. The full diagnostic plan is mainly a combination of a pelvic ultrasound besides blood tests of specific parameters that indicate the presence of PCOS. Since PCOS is a hard-to-diagnose widespread hormonal disorder, blood tests, symptoms, and other parameters with the help of a computer can form a new and easy method to diagnose it. Therefore, we had successfully built a high performing diagnostic model using MATLAB. The data was obtained from the website Kaggle, and the dataset is called Polycystic Ovary Syndrome. In this paper various machine algorithms were employed by utilizing seven classifiers. Results demonstrated that Linear Discriminant classifier exhibits the best performance in terms of accuracy, while in terms of sensitivity, the KNN classifier had the best result. Also, a comparison with four other research papers that exploited the same PCOS dataset was done in terms of implementation platforms, evaluation methods, classifiers, classes, accuracy, and precision of each classifier. Our research excelled among all in terms of accuracy and varied in precedence with precision. MATLAB had shown substantial results and a great model fitting embedded approaches, scoring a high accuracy and precision outcome compared to other studies. Other improvements on the overall PCOS prediction can involve employing preprocessed ultrasound images with the features presented in the dataset.**

Keywords*— Machine Learning, Polycystic Ovary Syndrome, Classification, Prediction, Diagnosis, Region of Interest.

I. Introduction

PCOS is the most common endocrine disorder in women of reproductive age [1]. PCOS is characterized by the ovaries production of an abnormal number of androgens which are male sex hormones that are normally present in women in small amounts. These androgens can cause more problems with the women's menstrual cycle and' are a reason of many PCOS features [2]. PCOS has many symptoms including irregular menstrual cycles, acne, heavy periods, excess hair growth, thickened and dark areas of skin, weight gain, pelvic pain, oily skin, and difficulty in getting pregnant. It is specified by hyperandrogenism, insulin resistance, anovulation where the ovary does not release an oocyte during the menstrual cycle, and neuroendocrine disruption [3-4].

PCOS diagnosis can be tricky, because not everyone with PCOS has polycystic ovaries (PCO), nor does everyone with ovarian cysts have PCOS, so the pelvic ultrasound as a stand-alone diagnosis is not sufficient [5]. The full diagnostic plan is mainly a combination of a pelvic ultrasound besides blood tests of specific parameters that indicate the presence of PCOS. PCOS can be diagnosed well in adults opposed to the diagnosis of adolescents where in this age group, the symptoms of the PCOS is overlapped with the characteristics of puberty. The diagnosis in adults is set by three distinct sets, one is determined by the National Institute of Health Consensus Statement which defines PCOS as menstrual irregularity and evidence of hyperandrogenism [6]. Another set is determined by the ESHRE/ASRM where the PCOS is defined as two of three features including anovulation or oligo-ovulation, hyperandrogenism, and polycystic ovaries by ultrasound [7]. The third and final set is defined by the Androgen Excess and PCOS Society, that diagnose PCOS as hyperandrogenism with ovarian dysfunction of polycystic ovaries [8]. Although the PCOS diagnosis is determined by one of the three sets, it is still difficult to certainly diagnose it, one reason according to Dr. Darche is there is no universal definition of the condition, "There are multiple expert-derived criteria for the syndrome, which means there is no universal diagnostic test or algorithm that doctors used to assess patients," she said. The other reason is that symptoms vary between women and does not affect them the same, this makes the diagnosis even more ambiguous for doctors. Furthermore, symptoms might not necessarily point to PCOS, but could be

related to other endocrine issues, obesity, and hypothyroidism [9].

Since PCOS is a hard-to-diagnose widespread hormonal disorder, blood tests, symptoms, and other parameters with the help of a computer can form a new and easy method to diagnose it. By collecting clinical data and building a model by writing algorithms, Machine Learning has shown its efficiency in the health sector when it comes to diagnosing diseases accurately [10].

Thus, in this paper, we had successfully built a high performing diagnostic model using MATLAB with the clinical data obtained from Kaggle Dataset. Using seven different classifiers, comparing between their accuracies, and comparing the results with three research papers of the same topic with the same dataset, and viewing their confusion matrices. By integrating supervised Machine Learning in PCOS diagnosis, the process of diagnosis will become easier, resulting in early treatment for women who are suffering from this disorder. The remaining of this paper is sectioned as follows; literature survey, methodology, results, discussion, conclusion, and references.

II. LITERATURE SURVEY

A recent study was conducted in [16] to detect PCOS using machine learning algorithms. Five data mining techniques (random forest, decision tree, Naive Bayes, logistic regression, and artificial neural network) were employed to determine the best model, which was the random forest that performed well achieving a 93.06% of accuracy, 92.66% of precision, 93.52% of sensitivity, and 92.59% of specificity.

Earlier studies were conducted on the same dataset used in this paper. One study was performed in [13] were the linear support vector machine (SVM) was used with a precision (93.665%) as well as high accuracy (91.6%) and recall (80.6%). Another study was done in [14] to detect PCOS were the Naïve Bayes and Adaboost techniques had shown a high accuracy of (87.72%).

Regarding the studies in [12] and [11], the best classifiers were KNN and logistic regression respectively with accuracies (86.58%) and (87.20%) respectively. It is noted that some researchers had done specific feature selection and data filtering that guarantee good prediction results. Some of them chose the features relying on how medically reliable they are and managed to achieve considerably high accuracy and precision values.

III. METHODOLOGY

A. Dataset Description

The data was obtained from the website Kaggle, which is known among Machine Learning practitioners for its scientific datasets. The dataset is called Polycystic Ovary Syndrome, collected from 10 hospitals across Kerala-India, published 2020, contains clinical and physical parameters of 541 women—364 were healthy and 177 were diagnosed with PCOS. The clinical parameters included metabolic, hormonal, and biochemical such as FSH, LH, Hb, TSH, AMH, Vit D3, PRG, etc. While physical parameters included weight, height, age, measurements of waist, hip, and the ratio between the two. TABLE I shows the attributes that were chosen for the diagnosis prediction for the machine learning algorithm, PCOS is the response data (positive=1, negative=0). The attributes varied between types as nominal, continuous, and ordinal.

TABLE I. CHOSEN ATTRIBUTES FOR THE ML ALGORITHM

NO	Attributes	NO	Attributes
1	PCOS (Response)	23	Cycle Length (days)
2	Age (Year)	24	Marriage Status (years)
3	Weight (kg)	25	Pregnant (Y/N)
4	Height (cm)	26	No. of Abortions
5	BMI	27	FSH (mIU/ml)
6	Blood Group	28	LH(mIU/mL)
7	Pulse Rate (BMP)	29	I beta-HCG (mIU/mL)
8	RR (Breaths/min)	30	II beta- HCG (mIU/mL)
9	Hb (g/dl)	31	FSH/LH
10	Cycle(R/I)	32	Hip (inch)
11	Cycle Length(days)	33	Waist (inch)
12	Waist: Hip Ratio	34	Pimples (Y/N)
13	TSH (mIU/L)	35	Fast Food (Y/N)
14	AMH (ng/mL)	36	Reg. Exercise (Y/N)
15	PRL (ng mL)	37	BP_Systolic (mmHg)
16	Vit D3 (ng/mL)	38	BP_Diastolic (mmHg)
17	PRG (ng/mL)	39	Follicle No.(L)
18	RBS (mg/dl)	40	Follicle No.(R)
19	Weight gain (Y/N)	41	Avg. F size (L) (mm)
20	Hair growth (Y/N)	42	Avg. F size (R) (mm)
21	Skin darkening (Y/N)	43	Endometrium (mm)
22	Hair loss (Y/N)		

B. Classifier Models

This part of the report explains the steps that were carried out for the PCOS prediction using supervised machine learning from the obtained dataset mentioned above. For our research, we employed seven classifiers by writing an algorithm in MATLAB. The classifiers were the following: K-Nearest Neighbor (KNN), Neural Network, Naïve Bayes, Support Vector Machine (SVM), Classification Tree, Logistic Regression, and Linear Discriminant. The default MATLAB parameters were set except for the KNN where we 9 nearest neighbors yielded the highest accuracy. Fig. 1 illustrates the workflow chart of the machine learning model.

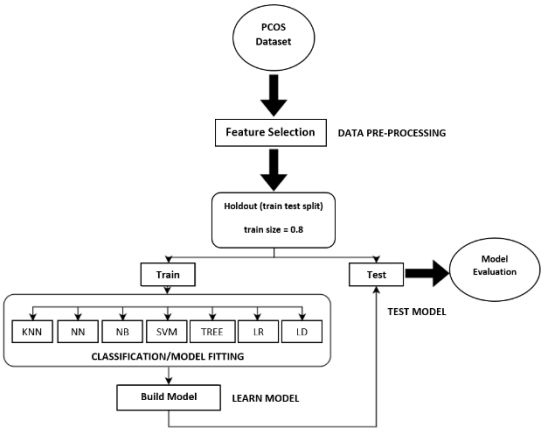

Fig. 1. Workflow Chart

C. *Feature selection and preprocessing*

This part of the report outlines the feature selection process and the preprocessing steps in writing the algorithm. Feature selection is a procedure that takes out irrelevant features to reduce overfitting and improve accuracy by only including informative and significant features that impact the diagnosis [11-14].

Input: PCOS training dataset.

Output: PCOS diagnosis (0/1).

Process:

1. Obtaining dataset.
2. Exclusion of serial number and patient file number features to increase accuracy of diagnosis due to their irrelevance.
3. Read the dataset table using readtable function.
4. Split the data into train and test data using cvpartition function (holdout method, 20% testing).
5. Fit the models using fitc function.
6. Predict the diagnosis using predict function.
7. Calculate the accuracy.
8. Visualize the performance of the model via the confusion matrices using confusionmat function.

IV. Results and Discussion

A. *Classifier performace comparison*

In this section, we will compare between the performances of each of the seven classifiers used in this research paper in terms of accuracy, precision, sensitivity, and specificity. After that, we will view the confusion matrices that were obtained for each classifier.

In this study, we calculated the performance parameter using the following equations: [15]

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

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

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

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

Where TP, TN, FP, FN are abbreviation for true positive, true negative, false positive, and false negative respectfully. The values of TP, TN, FP, and FN were obtained from the confusion matrices as shown in the fig. 2 below:

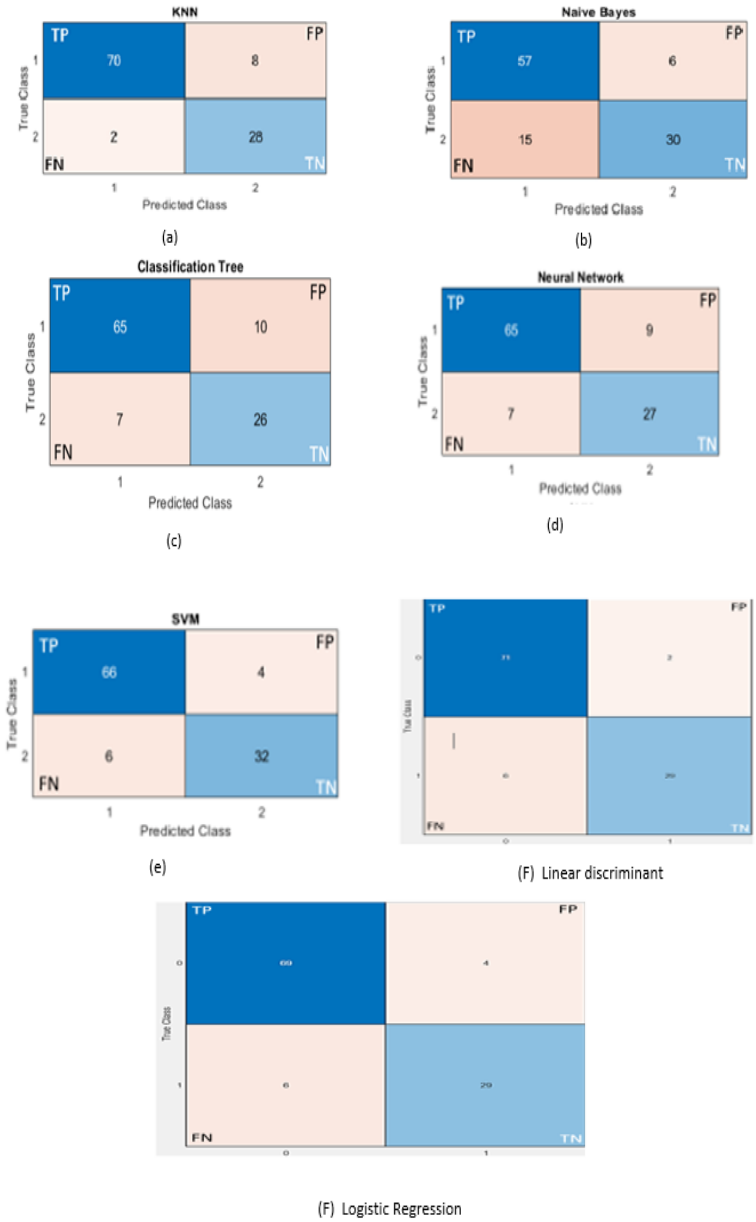

Fig. 2. Confusion Matrices for seven classifier

TABLE II. VALIDATION PARAMETERS

Classifier	Accuracy	Precision	Sensitivity	Specificity
KNN	90.74%	89.74%	97.22%	77.78%
NN	85.19%	87.84%	90.27%	75.00%
NB	80.56%	90.48%	79.16%	83.33%
SVM	90.74%	94.29%	91.67%	88.89%
Tree	84.26%	86.67%	90.27%	72.22%
Logistic	90.7%	94.52%	92.00%	87.87%
Linear	92.60%	97.6%	92.20%	93.55

The results have shown that in terms of accuracy, precision, and specificity—linear discriminant classifier was the most efficient. While in terms of sensitivity, the KNN classifier had the best result. TABLE II below shows the calculated parameters of validation for each model:

B. Comparison with literature

We compared between our research and four other research papers that used the same dataset in terms of accuracy and precision, and it was concluded that our model had shown a high-performing response prediction for PCOS. TABLE III below illustrates the comparison between results obtained from different experiments, while TABLE IV shows the accuracies and precision acquired by relevant papers. Our research excelled among all in terms of accuracy and varied in precedence with precision. Please note that in TABLE IV, for reference [11], we acquired the accuracy and precision values of 20% test size of the holdout method for comparison, and for reference [14], we compared between the accuracy and precision values of the Python platform only.

TABLE III. COMPARISON WITH LITERATURE BASED ON METHOD

Ref #	Dataset	Implementation Platform	Evaluation Method	Classifiers	Classes
11	All used the same PCOS dataset	Spyder Python	Hold out & cross validation	GB/RF/LR/RFLR	PCOS vs No PCOS
12		Spyder Python	Holdout	LR/LD/KNN/CART/RF/SVM	
13		Not mentioned	Holdout	KNN/LSVM/PSVM	
14		Python-Scikit &RapidMiner	Cross-validation	KNN/SVM/RF/NB/NN/Bagging/ Adaboost/Average	
This paper		MATLAB	Holdout	KNN/NN/NB/SVM/Logistic/ Linear	

TABLE IV. COMPARISON WITH LITERATURE BASED ON PERFORMANCE

Classifier	Ref#	Accuracy	Precision
NN	11	x	x
	12	x	x
	13	x	x
	14	78.63	79%
	This Paper	85.19%	87.84%
Linear	11	x	x
	12	x	x
	13	x	x
	14	x	x
	This Paper	92.60%	97.60%
Tree	11	x	x
	12	82.92%	74.29%
	13	x	x
	14	x	x
	This Paper	84.26%	86.67%
NB	11	x	x
	12	84.14%	82.14%
	13	x	x
	14	63.18%	75%
	This Paper	80.56%	90.67%
KNN	11	x	x
	12	86.58%	83.33%
	13	80.67%	99.09%
	14	72.27%	73%
	This Paper	90.74%	89.74%
Logistic	11	87.20%	78%
	12	85.36%	95.23%
	13	x	x
	14	x	x
	This Paper	90.70%	94.52%
SVM	11	x	x
	12	82.92%	100%
	13	91.60%	93.66%
	14	84.54%	85%
	This Paper	90.74%	94.29%

V. Conclusion

In this study, a labeled dataset was employed to build a high performing model to predict the presence or absence of PCOS. The prediction of the diagnosis has been carried out using seven classifiers and the evaluation of validation parameters has been done. MATLAB had shown satisfying results and a great model fitting embedded approaches, scoring a high accuracy and precision outcome, not to mention its high speed and easiness of use. Further improvements of our model can include a more careful choice of the algorithm parameters. Other improvements on the overall PCOS prediction can involve employing preprocessed ultrasound images and extracting significant and relevant features from them alongside with the features presented in the dataset. This leads to an increase in the number of samples and an enhancement of the prediction performance. On top of that, the presented paper can be a milestone for building a completed CAD system, starting from acquiring images, preprocessing them, segmenting the ovarian region as a ROI, extracting relevant features and making an accurate diagnosis based on them.

References

[1] Goodman, Neil F., et al. "American Association of Clinical Endocrinologists, American College of Endocrinology, and androgen excess and PCOS society disease state clinical review: guide to the best practices in the evaluation and treatment of polycystic ovary syndrome-part 1." Endocrine Practice 21.11 (2015): 1291-1300.

[2] Polycystic ovary syndrome (PCOS). Johns Hopkins Medicine. (n.d.). Retrieved December 25, 2021, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/polycystic-ovary-syndrome-pcos.

[3] U.S. Department of Health and Human Services. (n.d.). What are the symptoms of PCOS? Eunice Kennedy Shriver National Institute of Child Health and Human Development. Retrieved December 25, 2021, from https://www.nichd.nih.gov/health/topics/pcos/conditioninfo/symptoms

[4] Crespo, Raiane P., et al. "An update of genetic basis of PCOS pathogenesis." Archives of endocrinology and metabolism 62 (2018): 352-361.

[5] Frossing, Signe, et al. "Quantification of visceral adipose tissue in polycystic ovary syndrome: dual-energy X-ray absorptiometry versus magnetic resonance imaging." Acta Radiologica 59.1 (2018): 13-17.

[6] Zawadzski, J. K. "Diagnostic criteria for polycystic ovary syndrome: towards a rational approach." Polycystic ovary syndrome (1992): 39-50.

[7] Rotterdam ESHRE/ASRM‐Sponsored PCOS Consensus Workshop Group. "Revised 2003 consensus on diagnostic criteria and long‐term health risks related to polycystic ovary syndrome (PCOS)." Human reproduction 19.1 (2004): 41-47.

[8] Azziz, Ricardo, et al. "Criteria for defining polycystic ovary syndrome as a predominantly hyperandrogenic syndrome: an androgen excess society guideline." The Journal of Clinical Endocrinology & Metabolism 91.11 (2006): 4237-4245.

[9] Tian, Lifeng, et al. "Androgen receptor gene mutations in 258 Han Chinese patients with polycystic ovary syndrome." Experimental and Therapeutic Medicine 21.1 (2021): 1-1.

[10] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." Artificial Intelligence in medicine 23.1 (2001): 89-109.

[11] Bharati, Subrato, Prajoy Podder, and M. Rubaiyat Hossain Mondal. "Diagnosis of polycystic ovary syndrome using machine learning algorithms." 2020 IEEE Region 10 Symposium (TENSYMP). IEEE, 2020.

[12] Denny, Amsy, et al. "I-HOPE: detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques." TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE, 2019.

[13] Adla, Yasmine A. Abu, et al. "Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques." 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME). IEEE, 2021.

[14] Satish, CR Nandipati, XinYing Chew, and Khai Wah Khaw. "Polycystic Ovarian Syndrome (PCOS) classification and feature selection by machine learning techniques." (2020).

[15] Alqudah, Ali Mohammad, et al. "Employing image processing techniques and artificial intelligence for automated eye diagnosis using digital eye fundus images." Journal of Biomimetics, Biomaterials and Biomedical Engineering. Vol. 39. Trans Tech Publications Ltd, 2018.

[16] Marreiros, Marcello, et al. “Classification of Polycystic Ovary Syndrome Based on Correlation Weight Using Machine Learning”. Big Data Analytics and Artificial Intelligence in the Healthcare Industry, 2022.