Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms

Subrato Bharati[1], Prajoy Podder[2], M. Rubaiyat Hossain Mondal[3]
Institute of Information and Communication Technology
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
[1]subratobharati1@gmail.com, [2]prajoypodder@gmail.com, [3]rubaiyat97@yahoo.com

***Abstract*— This paper focuses on the data-driven diagnosis of polycystic ovary syndrome (PCOS) in women. For this, machine learning algorithms are applied to a dataset freely available in Kaggle repository. This dataset has 43 attributes of 541 women, among which 177 are patients of PCOS disease. Firstly, univariate feature selection algorithm is applied to find the best features that can predict PCOS. The ranking of the attributes is computed and it is found that the most important attribute is the ratio of Follicle-stimulating hormone (FSH) and Luteinizing hormone (LH). Next, holdout and cross validation methods are applied to the dataset to separate the training and testing data. A number of classifiers such as gradient boosting, random forest, logistic regression, and hybrid random forest and logistic regression (RFLR) are applied to the dataset. Results show that the first 10 highest ranked attributed are good enough to predict the PCOS disease. Results also demonstrate that RFLR exhibits the best testing accuracy of 91.01% and recall value of 90% using 40-fold cross validation applied to the 10 most important features. Hence, RFLR is suitable for reliably classifying PCOS patients.**

***Keywords*—** PCOS, gradient boosting, random forest, logistic regression, recall, precision.

I. Introduction

Polycystic ovary syndrome (PCOS) is a regular endocrine syndrome in females affected by elevated androgen. Treatment of PCOS is based on some cardinal features including anovulation, hyperandrogenism signs, and menstrual dysfunction. Some of the important symptoms of this disorder are pelvic pain, excess hair, acne, hirsutism velvety skin, male hormone, irregular periods, danker, etc. Approximately 5-10 % of reproductive age (15-49 years) women are suffered from this problem. The outbreak of PCOS was documented to be 8% and 4.8% in African American women and white women, respectively [1-2]. Women with this ovarian dysfunction are associated with hypertension, increased risk of cardiovascular disease, obesity, gynecological cancer, type 2 diabetes mellitus, etc. Moreover, recent research has shown that PCOS causes higher risk of first trimester miscarriage. PCOS causes in ovaries inappropriate growth of follicle that are prevented at a primary stage and miscarry to mature. This is one of the causes for infertility. Therefore, it is significant to screen the patients at a primary stage to prevent any serious moment of the PCOS disease. This paper focuses on the prediction of PCOS. The main contributions of this work as follows:

(i) selecting the most important attributes of PCOS patients for the given dataset using feature selection method.

(ii) applying machine learning algorithms on the important features of the PCOS dataset.

(iii) comparing the machine learning algorithms with those reported in the literature in terms of testing accuracy and recall.

The rest of the paper is organized as follows. Section II provides a brief literature review on the application of statistical tools and machine learning on the diagnosis of PCOS. Section III describes the overall methodology applied to conduct this research work. The process of finding the most significant features of PCOS is described in Section IV. The results obtained using the application of machine learning on PCOS dataset are described in Section V. Finally, Section VI provides the concluding remarks.

II. Literature Reviews

This section provides a brief literature work on PCOS [3-13]. Table I summarizes the literature survey. In the literature of PCOS, follicles detection has been described in two methods, one is stereology and the other is a sequential process of classification and feature extraction. In [3], polycystic ovary has been recognized conducting stereology to compute the number and the size of respective follicle and using Euclidean distance to measure the diameter of the follicle. On the other hand, in [4-5], Gabor Wavelet is used for feature extraction. Moreover, Conjugate Gradient-Fletcher Reeves and Lavenberg-Marquardt Optimization are used as a variation of backpropagation to classify PCOS [4-5]. In the work of [6], the segmentation of the follicles is done using particle swarm optimization (PSO) to make modifications to the fitness function. The authors in [7] conduct research using three classifiers namely (1) support vector machine (SVM) with RBF kernel, (2) k nearest neighbors (kNN) – Euclidean distance, and (3) neural network - learning vector quantization (LVQ). In [8], follicles are segmented using region growing scheme. This technique tests whether the neighbor of initial seeds should be added to segmentation region. The segmentation process of ultrasound image of PCOS has been done in [9, 10] using edge detection. The authors of [9] consider a median filter to remove noise from PCOS images. The main idea of this filter is to find a median in a specific picture element window. The center window will be updated with a median of the window. Meanwhile, Otsu global threshold [10] of the image is a way to find the pixel similarity to its neighbors. Otsu's thresholding method iterates through the threshold values. Furthermore, this threshold calculates a measure of the spread for the pixel levels on each side of the threshold. In the Canny edge detection, a computational approach to

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

Canny edge detection was presented in [11]. It can be used to detect follicles in the ultrasound image for the diagnosis of PCOS.

However, little work has been reported regarding the use of machine learning techniques on the clinical parameters for screening PCOS patients. So, this research focuses in the early detection of PCOS disease.

TABLE I. SUMMARY OF LITERATURE REVIEWS FOR PCOS

Ref	Methods used	Research objective
[3]	Euclidean distance to measure the diameter of the follicle	Determination of polycystic ovary syndrome
[4]	Modified backpropagation algorithm	Ultrasound images based PCOS detection
[5]	Wavelet-SVD	Digital image compression
[6]	Particle swarm optimization	PCOS detection
[7]	(1) SVM – RBF kernel, (2) KNN – Euclidean distance, (3) neural network - learning vector quantization (LVQ)	Follicle detection of PCOS using ultrasound images
[8]	Particle Swarm Optimization and Supervised Learning	PCOS classification using ultrasound images
[9]	Automated ovarian follicle recognition	PCOS image processing
[10]	Automated algorithm	Ovarian cysts detection

III. METHODOLOGY

This section briefly explains the experiments that are carried out to predict PCOS using machine learning classifiers. For this, a dataset with 43 attributes of 541 women are collected from the Kaggle repository [14]. Out of these 541 instances, 364 are for normal and the remaining 177 are for PCOS affected patients. For the experiments, Python programing language is used as a machine learning tool. For this, Anaconda distribution package, Scikit-learn library, Jupiter notebook, Spyder, Orange, etc. are used for the deployment of Python.

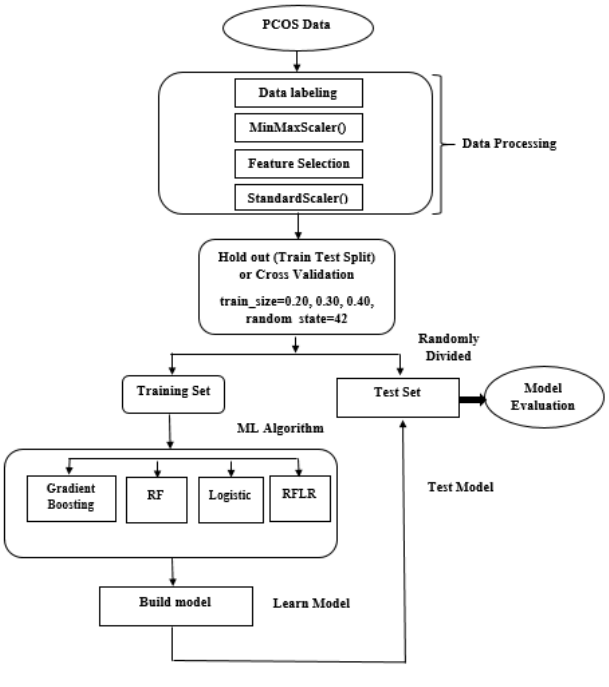

Fig.1. Work flow chart

The experiment has a number of stages. Fig. 1 illustrates the work flow process. Data labelling, encoding non-numeric columns into numeric, etc. are the stages of data pre-processing. Feature selection is applied after data pre-processing. Next, a number of classification algorithms including gradient boosting, random forest, logistic regression and hybrid random forest and logistic regression (RFLR) are used. Note that RFLR is implemented in Orange where random forest and logistic regression are aggregated using stacking process. The dataset is divided into training and testing portions using holdout and cross validation methods. For the holdout method, train_test_split() function from scikit-learn library is used in order to split the dataset. For separating the training and testing data samples, cross validation method is also applied using cross_val_score() function from scikit-learn library. These records are categorized into two classes: 0 and 1, where 0 means normal women and 1 means women having PCOS. At the data labelling stage, the non-numerical categorical attributes are converted into numerical values. The MinMax scaler or min-max normalization is used as follows:

$$p' = \frac{p-p_{min}}{p_{max}-p_{min}} \qquad (1)$$

Here, p denotes the original value of the feature, p_{min} is the minimum value of p, p_{max} is the maximum value of p, and p' is the normalized value of p. MinMax scaler mainly transforms features by scaling each feature to a range. Standard deviation of features is small due to scaling.

IV. FEATURE SELECTION

This section discusses the feature selection process. Feature selection is the procedure of taking some informative and significant features from a huge number of features. This can produce a better pattern characterization of multiple classes. Taking irrelevant features in the data can reduce the accuracy of the classification models. Feature selection can reduce overfitting and improve accuracy.

One form of feature selection approach is the filtering based univariate feature selection method which considers each feature independently that is with regard to the dependent variable. Each feature is scored individually on certain specified criteria and the features are then selected based on the higher scores or higher ranks. In this work, univariate feature selection method has been used for identifying the important features and finding the score. For selecting the *k* best features from the PCOS dataset, Algorithm 1 [15] is used.

Algorithm 1. Feature selection and other preprocessing

```
Input: A training dataset
Output: Selected features ranked according to p-value
Process:
1. Load the data using read_csv() function
2. Delete any null column using drop() function
3. Label the target using LabelEncoder() function
4. Divide the data and assign the dataset to x (without the target)
5. Assign the target to y
6. Scale the features for a range of 0 to 1 with MinMaxScaler() function
7. Obtain selected number of features using SelectKBest() function having arguments of chi2()
```

V. Results and Discussion

This section presents the results of the application of machine learning algorithms on the PCOS dataset. Using Algorithm 1, the ranking of all the 43 features are determined. Table II shows the ranking of the best 10 features. It can be seen that the ratio of Follicle-stimulating hormone (FSH) and Luteinizing hormone (LH) indicated as FSH/LH has the best ranking. The feature FSH and Anti-Mullerian hormone (AMH) have the second and the third best ranking, respectively. For finding the optimum number of features, different classifiers are applied to data for the case of different number of features. For 10 features with 20-fold cross validation, a number of classifiers are used. With this condition, the classification accuracy obtained by k nearest neighbors (kNN), random forest, decision tree, light GBM (LGBM), adaptive boosting (AdaBoost), gradient boosting, logistic regression, multilayer perceptron (MLP), hybrid random forest and logistic regression (RFLR) are 69.87%, 87.83%, 85.18%, 88.58%, 85.96%, 88.75%, 86.52%, 83.22% and 90.01%, respectively. It can be seen that good accuracy scores are obtained by gradient boosting, random forest, logistic regression, and RFLR, hence only these four classifiers are considered for the rest of the paper. Next, different number of features are considered for predicting PCOS patients. For the given dataset and for the four classifiers, the use of 10 or 14 number of features show competitive good results when both accuracy and low feature numbers are taken into consideration. This is because low number of features ensures low computation time and low complexity. Table III shows the values of precision, recall and testing accuracy for the four algorithms when 30-fold cross validation is used. It can be seen that the performance is almost the same for 10 and 14 features so the lower value that is 10 number of features are suitable for PCOS prediction. It can be seen that RFLR has the best classification accuracy being 90.01% and 89.27% for 10 and 14 features, respectively.

Next, the impact of cross validation number is investigated. Table IV shows that gradient boosting obtains its highest accuracy of 88.92% when 40 fold cross validation is applied. However, the same area under the curve (AUC) value of 84% is obtained at 10, 20, 30 and 40-fold cross validation. On the other hand, at 40-fold cross validation, random forest provides a good accuracy of 88.24%, recall value of 90% and AUC value of 90%. It can be seen that the performance of RFLR is the best at a cross validation of 40 where the accuracy, recall and AUC values are 91.01%, 90% and 92.90%, respectively.

Next, the effectiveness of holdout and cross validation methods is investigated for predicting PCOS. The comparison of these two approaches is performed for the four classifiers. Table V shows the performance results for different values of train test splitting. Test size of 20%, 30% and 40% of the total samples are considered. Note that a test size of 40% means, the training data is 60% of the total sample size. From Table V, it can be seen that RFLR obtains an accuracy of 88.52% when the test size is 40%. This value is lower than the value for RFLR with 40-fold cross validation as shown in Table IV. Gradient boosting performs the best when 40% is the test size where the accuracy value is 87.09% and AUC value is 86%. However, the accuracy value is lower and the AUC value is higher than the corresponding values obtained with 40-fold cross validation shown in Table IV. Table V also shows that random forest has accuracy, recall and AUC values of 88.94%, 89% and 88%, respectively when test size is 40%. For random forest, the value of accuracy obtained with holdout method shown in Table V is slightly higher than the value obtained with cross validation shown in Table IV. However, for random forest, the values of recall and AUC are higher for cross validation method than holdout method.

Table II. Ranking of best 10 features

Features name	Ranking
FSH/LH	1
FSH (mIU/mL)	2
AMH (ng/mL)	3
BMI	4
Weight gain (Y/N)	5
Follicle No. (L)	6
Follicle No. (R)	7
Avg. F size (L) (mm)	8
Cycle (R/I)	9
Cycle length (days)	10

Table III. Performance using 30-fold cross validation for different number of features

Classifier	Number of features	Precision	Recall	Accuracy
Gradient Boosting	10	85%	85%	88.72%
	14	85%	85%	88.72%
Random Forest	10	88%	88%	87.99%
	14	88%	88%	87.98%
Logistic regression	10	85%	85%	88.35%
	14	83%	83%	88.35%
RFLR	**10**	**89.90%**	**90%**	**90.01%**
	14	89.20%	89.30%	89.27%

Table IV. Performance using K-fold cross validation

Classifier	Value of K	AUC	Precision	Recall	Accuracy
Gradient Boosting	10	84%	85%	85%	87.43%
	20	84%	85%	85%	88.72%
	30	84%	85%	85%	88.72%
	40	84%	85%	85%	88.92%
Random Forest	10	91%	91%	90%	88.16%
	20	90%	90%	90%	87.79%
	30	89%	88%	88%	87.99%
	40	90%	90%	90%	88.24%
Logistic regression	10	82%	83%	83%	88.53%
	20	82%	83%	83%	88.53%
	30	82%	83%	83%	88.35%
	40	82%	83%	83%	88.39%
RFLR	10	92.90%	90.10%	90.20%	90.20%
	20	92.90%	89.90%	90%	90.01%
	30	92.90%	89.90%	90%	90.01%
	40	**92.90%**	**89.90%**	**90%**	**91.01%**

TABLE V. PERFORMANCE USING HOLDOUT METHOD

Classifier	Test size	AUC	Precision	Recall	Accuracy
Gradient Boosting	20%	78%	81%	82%	81.70%
	30%	85%	86%	87%	86.50%
	40%	86%	87%	87%	87.09%
Random Forest	20%	83%	85%	85%	85.30%
	30%	86%	87%	87%	87.11%
	40%	88%	89%	89%	88.94%
Logistic regression	20%	85%	87%	87%	87.20%
	30%	85%	86%	87%	86.50%
	40%	87%	88%	88%	88.47%
RFLR	20%	91%	87%	87%	86.96%
	30%	92%	87.5%	87%	87.60%
	40%	**93%**	**88.4%**	**88%**	**88.52%**

The performance of RFLR algorithm in classifying PCOS patients is comparable with the results reported in the literature. In this paper for RFLR and other algorithms, only 10 features are used which is less than that of the previous research papers. Therefore, the classification of the proposed RFLR and other methods used in this work require less computation time compared to the existing research works. The accuracy values of the studies in [4] and [7] are reported to be 93.92% and 82.55%, respectively. However, the studies in [3-10] use ultrasound images instead of the data/parameters used in this study. Therefore, the results in [3-10] cannot be directly compared with the result of this work. The work in [12] uses competitive neural network to obtain prediction accuracies of 78%, 80% and 80.84% for 16, 24, 32 PCOS features, respectively. In [13], Elman neural network is used and the obtained accuracy values are 68.8%, 66.7% and 78.1% for 16, 24 and 32 features, respectively. The work in [16] uses 12 features and exhibits an accuracy of 89.02%, but a low sensitivity or recall value of 74.19% using random forest classifier. On the other hand, we have already shown in this work that RFLR provides a high accuracy of 91.01% and a high recall value of 90% when 40-fold cross validation is applied.

VI. CONCLUSION

This paper presents data-driven diagnosis of PCOS disease in women. For this, the concepts of feature selection and machine learning algorithms are applied. A dataset of 541 patients obtained from Kaggle repository is used. Our results show that FSH/LH is the most important attribute among the 43 attributes in the dataset. Results indicate that good accuracy with lower computation time can be obtained when the best 10 features are used. A number of different classifiers are used on these 10 features. It is shown in this paper that gradient boosting, random forest, logistic regression and RFLR exhibit good accuracy and recall values. RFLR has the best testing accuracy of 91.01% and recall value of 90% when 40-fold cross validation is used to split the data into testing and training portions. The results of RFLR are comparable with the results reported in the literature. Furthermore, only 10 features are used in this paper which is less than that of the previous research papers leading to less computation time.

As a future work, the results obtained from this paper can be validated with a number of different datasets. Furthermore, a number of hybrid methods can be developed to increase the classification accuracy of the machine learning algorithms.

REFERENCES

[1] R. Pasquali et al., "PCOS Forum: Research in Polycystic Ovary Syndrome Today and Tomorrow", *Clin Endocrinol (Oxf)*, vol.74, no. 4, pp. 424–433, 2011. doi: 10.1111/j.1365-2265.2010.03956.x

[2] A. S. Laganà, S.G. Vitale, M. Noventa, and A. Vitagliano, "Current Management of Polycystic Ovary Syndrome: From Bench to Bedside", *International Journal of Endocrinology*, 2018. doi: 10.1155/2018/7234543

[3] Adiwijaya, B. Purnama, A. Hasyim, M. D. Septiani, U. N. Wisesty and W. Astuti, "Follicle detection on the usg images to support determination of polycystic ovary syndrome", *3rd International Conference on Science & Engineering in Mathematics, Chemistry and Physics 2015 (ScieTech 2015)*, vol. 622, 2015.

[4] U. N. Wisesty, J. Nasri and Adiwijaya, "Modified backpropagation algorithm for polycystic ovary syndrome detection based on ultrasound images", *Recent Advances on Soft Computing and Data Mining - The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia*, pp.144-151, August 18-20, 2016.

[5] Adiwijaya, M. Maharani, B. Dewi, F. Yulianto and B. Purnama, "Digital image compression using graph coloring quantization based on wavelet-svd", *2013 International Conference on Science & Engineering in Mathematics, Chemistry and Physics (ScieTech 2013)*, vol. 423, 2013.

[6] E. Setiawati, Adiwijaya and A. Tjokorda, "Particle swarm optimization on follicles segmentation to support pcos detection*", 3rd International Conference on Information and Communication Technology (ICoICT)*, pp 369-374, 2015.

[7] B. Purnama, U. N. Wisesti, Adiwijaya, F. Nhita, A. Gayatri and T. Mutiah, "A classification of polycystic ovary syndrome based on follicle detection of ultrasound images", *3rd International Conference on (IEEE) Information and Communication Technology (ICoICT)*, pp 396-401, 2015.

[8] E. Setiawati, Adiwijaya, T. A. B Wirayuda, W. Astuti, "A Classification of Polycystic Ovary Syndrome Based on Ultrasound Images Using Supervised Learning and Particle Swarm Optimization", *Advanced Science Letters* , vol. 22, pp.1997-2001, 2016.

[9] P. Mehrotra, C. Chakraborty, B. Ghoshdastidar, S. Ghoshdastidar and K. Ghoshdastidar, "Automated ovarian follicle recognition for polycystic ovary syndrome", *International Conference on Image Information Processing (ICIIP)*, pp 1-4, 2011.

[10] S. Rihana, H. Moussallem, C. Skaf and C. Yaacoub , "Automated algorithm for ovarian cysts detection in ultrasonogram", *2nd International Conference on Advances in Biomedical Engineering (ICABME)*, pp 219-222, 2013.

[11] J. Canny, "A Computational Approach to Edge Detection", *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 679-698, 1986. doi:10.1109/TPAMI.1986.4767851

[12] R. M. Dewi, Adiwijaya, U. N. Wisesty and Jondri, "Classification of polycystic ovary based on ultrasound images using competitive neural network", *International Conference on Data and Information Science*, vol. 971, 2017.

[13] I. F. Thufailah, Adiwijaya, U. N. Wisesty and Jondri, "An implementation of Elman neural network for polycystic ovary classification based on ultrasound images", *International Conference on Data and Information Science*, vol. 971, 2017.

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

[15] M. Raihan-Al-Masud and M. R. H. Mondal, "Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms". *PLOS ONE*, vol. 15, no. 2: e0228422, 2020. https://doi.org/10.1371/journal.pone.0228422

[16] A. Denny, A. Raj, A. Ashok, C. M. Ram and R. George, "i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques," *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Kochi, India, 2019, pp. 673-678.