Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques

Palak Mehrotra,Jyotirmoy Chatterjee,Chandan Chakraborty
School of Medical Science and Technology
Indian Institute of Technology
Kharagpur, India
chandanc@smst.iitkgp.ernet.in, palakmehrotra@gmail.com

BiswanathGhoshdastidar,Sudarshan Ghoshdastidar
G D Institute for Fertility Research
Kolkata, India
bgdastidar@gmail.com, sudarsan.ivf@gmail.com

Abstract— **Polycystic Ovary Syndrome (PCOS) is one of the most common type of endocrine disorder in reproductive age women. This may result in infertility and anovulation. The diagnostic criterion includes the clinical and metabolic parameters which act as an early marker for the disease. We described a method that automates the PCOS detection based on these markers. Our algorithm involves the formulation of feature vector based on the clinical and metabolic features and statistically significant features for discriminating between normal and PCOS groups are selected based on two sample t-test. To classify the selected feature Bayesian and Logistic Regression (LR) classifier are used. An automated system will act as an assisted tool for the doctor for saving considerable time in examining the patients and hence reducing the delay in diagnosing the risk of PCOS. The study demonstrated that the performance of Bayesian classifier is better than the logistic regression. The overall accuracy of Bayesian classifier is 93.93% as compared with logistic regression i.e. 91.04%.**

Keywords-Polycystic Ovary Syndrome; Logistic Regression; Bayesian Classifier;

I. INTRODUCTION

Polycystic ovary syndrome (PCOS) is a heterogeneous endocrine disorder affecting most women of reproductive age [1]. This syndrome was initially described by Stein and Leventhal in 1935[2]. Nearly 5-10 % of reproductive age women are affected with this abnormality [3]. The prevalence of PCOS was reported to be 4.8% and 8% in white women and African American women respectively [4]. In Spain, the figure is 6.8% and 13% of Mexican American women have been reported [5]. Women with this ovarian dysfunction are associated with increased risk of cardiovascular disease, type 2 diabetes mellitus, obesity, hypertension, gynecological cancer [7-12]. Moreover, recent studies have shown greater risk of first Trimester miscarriage [13]. The symptoms associated PCOS include obesity, irregular menstrual cycle, excessive production of male hormone, acne, hirsutism [14]. PCOS leads to inappropriate development of follicle in the ovaries which are arrested at an early stage and fail to mature [15]. This is one of the reasons for infertility [16]. It is therefore important to screen the patients at an early stage to prevent any serious consequence of the disease.

The automated screening systems for PCOS disease based on ultrasound image have been reported in the literature [17]. But no literatures have been recorded regarding the use of machine learning techniques based on clinical parameters for screening of PCOS patients. So this automated screening system will help in the early detection of the disease.

The present study aims at statistically evaluating metabolic and clinical features on the basis of probability density function and box plot. Thereafter, the statistical parameters were fed to the two statistical models i.e. logistic regression and Bayesian classifier for the Polycystic Ovary disease prediction. Both the method were applied on the same female population. Then a comparative evaluation was done between two methods for their accuracy in calculating the disease probability.

The organization of paper is as follows: section 2 describes the materials and methods; section 3 discusses the results obtained by the method. Some discussions are present in section 4 and finally section 5 is the conclusion.

II. MATERIALS AND METHODS

A. Data Collection

The study was conducted on the patients coming to Ghosh Dastidar Institute for Fertility Research (GDIFR), Kolkata between March 2010 and April 2011. A pre-designed form provided by the clinic was used to record the medical history and other physical examination of each patient. Of these 250 women included in this study 150 were having polycystic ovary (PCO) and the rest 100 were normal. The American Society for Reproductive Medicine (ASRM) and European Society of Human Reproduction and Embryology (ESHRE) jointly established the diagnostic criteria for PCOS. The three criteria include: oligo-and anovulation (failure to ovulate), clinical and/or biochemical sign of excessive production of male hormone and the presence of polycysts in at least one of the ovaries (ultrasound examination). The confirmation of PCO should be based if the patient fulfills any two criterions as recommended by ASRM/ESHRE [6] The diagnosis of PCOS was made based on criterion: (1) Cycle length (2) clinical and metabolic feature (3) polycystic ovaries (presence of 12 or more follicles measuring 2-9 mm in diameter or

increased ovarian volume) [18] with the exclusion of hyperthyroidism and Cushing syndrome.

Transvaginal ultrasound was performed in all the patients coming to the clinic. The ultrasound was performed with a 7 MHz transducer (General Electricals, Milwaukee, USA) using standard protocol. The virgin or refusing patients, as well as those patients where no follicle was found in either of the ovaries were excluded from the analysis.

The block diagram for the automated screening of PCOS patients is shown in figure 1.

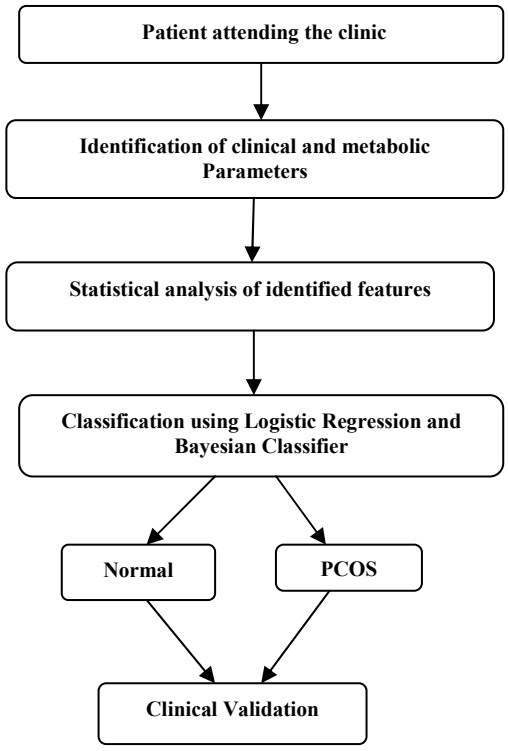

Figure 1: Block diagram of the methodology.

B. Feature Identification

The data collected from the patient was used with emphasis on menstrual cycle, metabolic and clinical features. Menstrual cycles were defined as (1) regular cycle i.e. a cycle with an interval of 21 to 35 days; (2) oligomenorrhea i.e. an interval of 36 days or longer; (3) infrequent i.e. interval of less than 35 days; (4) Irregular i.e. shorter than 21 days [5].

The body mass index (BMI) was calculated as weight in kilograms divided by height in meters. The patients having (BMI≤24) considered as normal weight while patients having weight (BMI≥25) considered as overweight based on World Health Organization (WHO). The fasting blood sugar was obtained in the morning time. The level of follicle stimulating hormone (FSH) and Luteinizing hormone (LH) were measured by hormonal kit. Women having polycystic ovaries generally show abnormal level of FSH and LH [19].

C. Feature Analysis

The main idea behind feature selection in machine learning is to build a subset of significant feature which is used for training the classifier and improve its comprehensibility. The clinical and metabolic parameters such as Age, BMI, LH, FSH, Systolic Blood Pressure, Diastolic Blood Pressure, Cycle Length, Fasting Blood Sugar, and Post Prandial Blood Sugar are taken for input patients. Statistical analysis is done using two sample t-tests to select a significant subset of original feature. The features having p value <0.001 are considered significant. Also the kernel density and box plot was used for cross validation.

D. Multivariate Logistic Regression

The logistic regression model is used for dichotomous categorical outcomes like the case as PCO and normal [22]. In such cases, multivariate logistic regression provides the best prediction possible based on the dependent variable [23]. Let us assume Y as dependant variable , so

$$Y = \begin{cases} 1 \text{ if the patient has PCOS.} \\ 0 \text{ if the patient is Normal.} \end{cases}$$

Where $\underline{x}^* = \{x_1, x_2, x_3, x_4\}$ are the significant parameters as predicted by *t*-test. The multivariate logistic regression was fitted to calculate the disease probability of the patient based on the significant parameters. The logistic model is as follows

$$\Pi(x) = P\left(Y = 1 \middle| \underline{x}^* \right) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{i=1}^{n} \underline{\beta_i} \underline{x_i} \right)}} \quad (1)$$

$\Pi(x)$ represents the probability of having the Polycystic Ovary Syndrome. The Maximum likelihood Method is used to estimate the regression coefficients $\beta_1, \beta_2, \beta_3 \ldots\ldots\ldots \beta_n$. A threshold value ' α ' is assumed according to the decision maker for the formulation of the decision rule as rule as:

$$P\left(Y = 1 \middle| \underline{x}^* \right) \geq \alpha \text{ classify in PCOS}$$

$$P\left(Y = 1 \middle| \underline{x}^* \right) \leq \alpha \text{ classify in Normal.}$$

E. Bayesian Classifier

The Bayesian classifier is the most useful and efficient probabilistic learning technique in the field approach machine learning [24]. It is a supervised statistical method used for binary and multiclass classification. This is based on Baye's theorem named after Thomas bayes [25]. Moreover it can be used for solving diagnostic and predictive problem. In this suppose we have an n-dimensional feature space with values

$$X = \{x_1, x_2, x_3 \ldots\ldots\ldots\ldots x_n\}$$

For this data we have n = 4. In Bayesian classifier which assigns the given data values into m classes like $\beta_1, \beta_2, \beta_3 \ldots\ldots\ldots \beta_m$. Then according to the greater posterior probability the new data will be classified in any of the classes. That means, X will be classified to class B_i

if $P(B_i | X) > P(B_j | X)$ for all i and j such that $1 \leq j \leq m, j \neq i$

The Bayes formula is expressed as

$$P\left(B_i \left| \underline{x}^* \right.\right) = \frac{P(B_i) \times P\left(\underline{x}^* \left| B_i \right.\right)}{\sum_{i=1}^{n} P(B_i) \times P\left(\underline{x}^* \left| B_i \right.\right)} \quad (2)$$

Since $P\left(B_i \left| \underline{x}^* \right.\right)$ is the posterior probability (or posterior) is the probability of the state of nature being B_i given that feature value of X has already being measured. We say $P\left(\underline{x}^* \left| B_i \right.\right)$ as the likelihood function or class conditional probability of B_i with respect to X. The $P(B_i)$ is the prior probability that shows the initial probability value originally obtained before any additional information.

III. RESULTS AND DISCUSSION

A total of 200 patients were considered for the study among which (*n=50*) patients were considered as normal while rest (*n = 150*) belong to PCOS group. The 9 parameters which were considered are listed in The Table 1. The statistically significant features are used further for automated classification using Bayesian Classifier and Logistic Regression into normal and PCO group.

Table 1: Summary statistics and statistical significance of features for normal and PCOS group

Parameter analyzed	Normal (n= 50)	Abnormal (n= 150)	*t*-value	*p*-value
Age (years)	32.24 ± 2.02	31.24 ± 2.48	1.520	0.134*
BMI (kg/m^2)	22.86 ± 2.05	26.94 ± 2.99	-8.891	<0.001
Basal FSH Level (IU/L)	5.77 ± 1.21	4.61 ± 1.85	-3.935	<0.001
Basal LH Level (IU/ L)	5.79 ± 3.16	10.10 ± 7.00	-3.816	<0.001
Systolic B.P (mm Hg)	126.72 ± 9.27	130 ± 13.83	-0.866	0.390*
Diastolic B.P (mm Hg)	78.40 ± 7.38	83.84 ± 7.81	-3.059	0.013*
Cycle Length (days)	28.66 ± 4.04	37.64 ± 4.51	-11.446	<0.001
Fasting Blood Sugar (mg/dL)	83.76 ± 7.04	90.48 ± 8.24	-0.256	0.799*
Post Prandial Blood Sugar (mg/dL)	88.26 ± 11.27	104.97 ± 24.71	-4.391	0.022*

In Table 1 the parameters having *p*-value < 0.001 were considered significant. In the Figure 3 below Box-plot [20] shows the significant variability of the medians for normal and PCOS group. We can easily see significant difference in the median as the notches of both the groups do not overlap. Moreover, the ranges for both the significant parameters are also variable which validate the results obtained by the t test. The kernel density functions [21] showed Figure 2 shows the very less overlapping between the two groups this cross verifies the results.

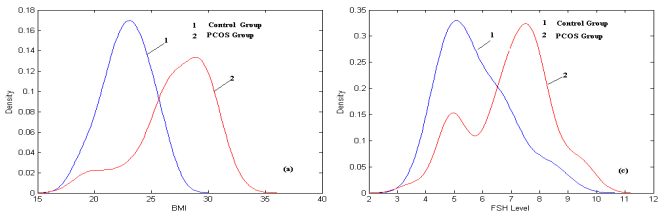

Figure 2: Kernel density plot for normal and PCOS group for (a) Body Mass Index (BMI), (b) FSH Level

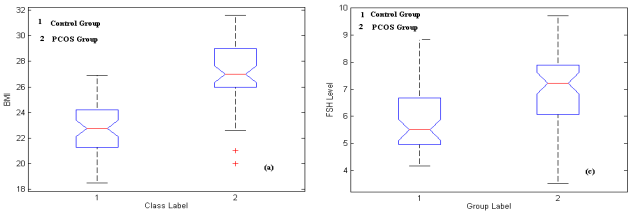

Figure 3: Box-Whisker's plot for normal and PCOS group for Body Mass Index (BMI), (b) FSH Level

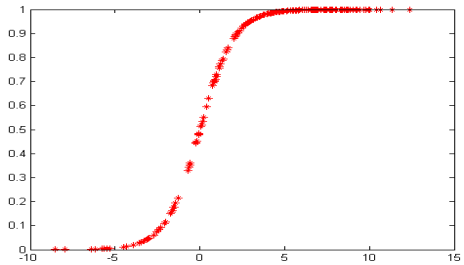

Figure 4: Multivariate logistic regression curve

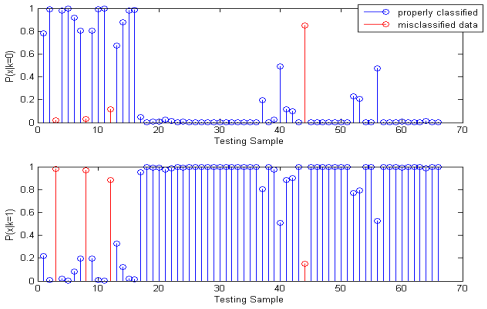

Figure5: Posterior probability for normal and PCOS group for Bayesian classifier

Table 2: Confusion matrix for Bayesian classifier

	Normal	PCOS
Normal	13	3
PCOS	1	49

Table 3: Confusion matrix for logistic regression

	Normal	PCOS
Normal	14	3
PCOS	3	47

Table 4: Comparison of accuracy for Bayesian and Logistic Regression

Classifiers	Sensitivity	Specificity	PPV	Accuracy
Logistic Regression	82.36 %	94 %	82.35 %	91.04 %
Bayesian Classifier	92.85 %	94.23 %	81.27 %	93.93 %

The fitted model as described in equation (1) the value of Z = -32.9656 + 0.3497*x_1 + 0.1258*x_2 + 0.5351*x_3 + 0.5435*x_4. If an undiagnosed patient having feature vector as $\underline{x}^*$ = (7.97, 3.3, 27, 30), then disease probability

$$P\left(y=1\left|\underline{x}^*\right.\right)=\frac{1}{1+e^{-(0.9890)}}=0.7290 \qquad (3)$$

According to rule it can be inferred that patient having 72.90 % risk of having PCOS. If the threshold value is set to 0.5 then the values below this are considered as normal group while values above this are considered high risk of having PCOS. The Figure 4 graph shows the relation between the Z value and the disease probability for the patient data. Considering the significant parameters for Bayesian classification 3 fold cross validation was done to find the accuracy.

Similarly, in the Bayesian classification if similar patient having feature vector $\underline{x}^*$ = [7.97, 3.3, 27, 30]. So the posterior probability of the patient falling in PCOS and normal class respectively

$$\mathrm{P}\left(B_1\left|\underline{x}^*\right.\right)=\frac{P(\mathrm{B}_1)\mathrm{P}\left(\underline{x}^*\left|\mathrm{B}_1\right.\right)}{\mathrm{P}(\mathrm{B}_1)\times\mathrm{P}\left(\underline{x}*\left|\mathrm{B}_1\right.\right)+\mathrm{P}(B_0)\times\mathrm{P}\left(\underline{x}^*\left|\mathrm{B}_0\right.\right)}=0.8389$$

$$\mathrm{P}\left(\mathrm{B}_0\left|\underline{x}^*\right.\right)=\frac{P(\mathrm{B}_0)\mathrm{P}\left(\underline{x}^*\left|B_0\right.\right)}{\mathrm{P}(\mathrm{B}_1)\times\mathrm{P}\left(\underline{x}*\left|\mathrm{B}_1\right.\right)+\mathrm{P}(B_0)\times\mathrm{P}\left(\underline{x}^*\left|\mathrm{B}_0\right.\right)}=0.1611$$

If the $\mathrm{P}\left(\mathrm{B}_1\left|\underline{x}^*\right.\right)>\mathrm{P}\left(\mathrm{B}_0\left|\underline{x}^*\right.\right)$ then the patient will fall in PCOS group else normal group. The data was divided into training and testing data. The confusion matrix was computed for the testing data as shown in Table 2. In Figure 5 the posterior probability of the test sample for both the classes is being plotted against the sample index. It shows the comparison between the posterior probabilities of the two classes. The data marked in red are the misclassified data from the testing sample. The same testing data was fed to the regression model and the confusion matrix was computed as shown in Table 3.

Concerning the statistically significant parameters, both the statistical approaches showed that the FSH, LH, BMI and cycle length as the most discriminating features. The influences of these parameters are extensively studied in literature and authors have confirmed their importance [19, 26].

The main aim of this paper was to review two classifiers: logistic regression and Bayesian for the screening of patients into normal and PCOS group. The sensitivity, specificity, positive predictive value (PPV) and overall accuracy of the classifiers is being summarized in the Table 4.

From the results, both the classifiers have achieved reasonable performance but both have some drawback.

The Bayesian classifier is a supervised classification technique and a good approach for classification technique. Moreover Bayesian is very easy and simple to understand and manual computation is less complex.

On the other hand logistic regression works well when the data set is large. Moreover the logistic regression requires higher computational labors than Bayesian classifier.

Using this dataset we obtained best result from Bayesian classifier (93.93%) as compared to Logistic regression (91.04%). This study suggests that Bayesian classifier is a good option over logistic regression in the early screening of PCOS patients.

IV. CONCLUSION

In this paper a new approach for PCOS screening is proposed. The two classifiers used provided consistent result in all folds. Based on four clinical features: FSH, LH, BMI, Cycle Length the occurrence probability for PCOS was predicted. The Bayesian classifier gives higher accuracy than the logistic regression. The use of probabilistic model will help the doctors for the early screening of the patients who are more likely to develop the disease. There is a need to improve this accuracy by the use of other classifiers.

ACKNOWLEDGEMENT

The author would like to thank Sudarshan Ghoshdastidar and Biswanath Ghoshdastidar, GDIFR, India and J.Chatterjee for their valuable clinical support and guidance.

REFERENCES

[1] Gautam N. Allahbadia , Rubina Merchant, "Polycystic ovary syndrome and impact on health," Middle East Fertility Society Journal, vol.16, pp. 19–37, october 2010.

[2] Morva Tahmasbi Rad," BMI role in treatment of infertile patients with polycystic ovary syndrome," International Congress Series , vol.1271, pp.34-37 , 2004.

[3] S.franks, "Medical progress: polycystic ovary syndrome," New England Journal of Medicine, vol.333(13):853-861, 1995.

[4] Asa Lindholm , Liselott Andersson , Mats Eliason , Marie Bixo,Inger Sundstom-Poromaa,"Prevalence of Symptoms associated with Polycystic Ovary Syndrome," International Journal of Gynecology and Obstetrics, vol.102, pp.39-43, january 2008.

[5] Lin Li, Dongzi Yang, Xiaoli Chen, Yaxiao Chen, Shuving Feng, Liangan Wang,"Clinical and Metabolic feature of Polycystic Ovary Syndrome," International Journal of Gynecology and Obstetrics, vol.97, pp.129-134, january 2007.

[6] A . H. B elen , J.S.E.Laven ,S-L Tan , D. Dewailly , "Ultrasound Assessment of the Polycystic Ovary : intenational Consesus Defination," Human Reproduction Updates,vol. 6, pp.505-514, 2003.

[7] I.F. Stein, M.L. Leventhal, "Amenorrhea associated with bilateral polycystic ovaries," American Journal. Obstetrics and. Gynecology, 1935; 29; 181-191.

[8] Ricardo Azziz , Keslie S. woods, Rosario Reyna , Timothy J. Key , Eric S. Knochenhauer , Bulent O. Yildir,"The Prevalence of the Polycystic Ovary Syndrome in an Unselected Population," The Journal of Clinical Endocrinology and Metabolism,vol.89(6), pp.2745-2749, 2004.

[9] Ovalle F , Azziz R. "Insulin Resistance, polycystic oavry syndrome and Type 2 diabetes mellitus," Fertility and Sterility , vol.77, pp1095-1105, 2002.

[10] Wild RA , "Long term health consequence of PCOS," Human Reproduction Updates, vol.24, pp231-241 , 2002.

[11] Legro RS," Polycystic Ovary Syndrome and Cardiovascular Disease : premature association?",Endocrine Reviews,vol24, pp.302-312, 2003.

[12] P Hardiman , OS Pillay, W Atiomo ,"Ploycystic Ovary Syndrome and Endometrial Carcinoma," Lancet, vol.361, pp.1810-1812 , 2003.

[13] Richard S. Legro, "Polycystic ovary syndrome: the new millennium," Molecular and Cellular Endocrinology, vol.184, pp.87-93, 2001.

[14] Maryruth J. Lawrence , Mark G. Eramian , Roger A. Pierson , Eric Neufeld, "Computer Assisted Detection of Polycystic Ovary Morphology in Ultrasound Images," IEEE Transl. Computer Society, 2007[Fourth Canadian Conference on Computer and Robot Vision , 2007

[15] Wafaa M Aboul Enien, Nadia A Barghash, Fayrouz S Mohamed Ali, Clinical, "Ultrasonographic and Endocrine predictors of Ovarian response to Clomiphene Citrate in Normogonadotropic Anovulatory Infertility," Middle East Fertility Society Journal,vol.9, 2004.

[16] Gerard S Convay , " Polycystic Ovary Syndrome: Clinical Aspects," Baillière's Clinical Endocrinology and Metabolism , vol. 10 , 1996.

[17] Yinhui Deng, Yuanyuan Wang, Ping Chen," Automated Detection of Polycystic Ovary Syndrome from Ultrasound Images," IEEE Transl. , 30th Annual International IEEE EMBS Conference,2008.

[18] S. Jonard , Y. Robert, C. Cortet-Rudelli , P. Pigny , C. Decanter, D. Dewailly , "Ultrasound Examination of Polycystic Ovaries: is it worth counting the follicle," Human Reproduction,Vol.18,pp598-603,2003.

[19] Joop S.E. Lavena, Annemarie G.M.G.J. Muldersa, Evert J.P. van Santbrinka, Marinus J.C. Eijkemansb, Bart C.J.M Fauser, "PCOS: Backgrounds, evidence and problems in diagnosing the syndrome," International Congress Series,vol.1279,pp.10-15,2005.

[20] L.S. Nelson, "Evaluating Overlapping Confidence Intervals," Journal of Quality Technology, vol.21, pp. 140-141, 1989.

[21] S.Towers, "Kernel probability density estimation methods," Proceedings of the Advanced Statistical Techniques in Particle Physics, pp.107-111, 2002.

[22] Joseph J Guido, Paul C Winters, Adam B Rains, Logistic Regression Basics, NSEGU, 2006.

[23] Gaafar M. Malik, Mostafa A. Abolfotouh, Suliman Jastania, Nader Morad, Eldawi N. Eltayeb, Gulam Saydain, "A Logistic Regression Model to Predict Nodal Malignancy among Cases with Lymphaedenopathy," Annals of Saudi Medicine, Vol 18, No 6, 1998.

[24] D. Testi , A. Cappello,L.Chiari , M. Viceconti , S.Gnudi , "Comparision of Logistic and Bayesian Classifier for Evaluating the Risk of Femoral Neck Fracture in Osteoporotic Patients," Medical and Biological Engineering and Computing,vol.39, pp.633-637, 2001.

[25] Duda, R.O. and Hart, P.E. (1973). Pattern Classification and Scene Analysis. New York: John Wiley and Sons .

[26] Stephen Franks, "Diagnosing polycystic ovary syndrome," Elsevier, 2006.