Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome

Malik Mubasher Hassan
BGSB University
Rajouri, J&K-India

Tabasum Mirza
Dept. of School Education
Govt. of J&K-India

ABSTRACT
Artificial intelligence can be used in healthcare systems for diagnostic purposes to handle large amounts of clinical data with much accuracy and precision. One of the commonest health issue found in the young women is Polycystic Ovarian Syndrome (PCOS) and it is basically a complex health disorder affecting women of reproductive age group that can be diagnosed based on clinical symptoms like increased body mass index, elevated hormone levels, hair loss, acne, skin darkening, hirustism, cycle length, endometrial thickness, high blood pressure levels, etc. Correct diagnosis is the baseline of any proper treatment and in this research paper we are using machine learning approaches like Support Vector Machine, CART, Naive Bayes Classification, Random Forest and Logistic Regression to diagnose PCOS based on the clinical data of patients. The results were analyzed and performance of the algorithms was validated on the basis of accuracy, precision, recall, F-statistics, and Kappa Coefficient. The validation metrics indicate the highest i.e. 96% accuracy of the Random Forest algorithm in the diagnosis of PCOS on giving data.

Keywords
Polycystic Ovarian Syndrome, Machine Learning, Diagnosis and Random Forest

1. INTRODUCTION
Polycystic Ovarian Syndrome also called as Stein–Leventhal syndrome is an endocrine disorder affecting 5 to 10 percent of women in reproductive age (12-45 years) [1].The disease was first discovered by Irving F. Stein, Sr., and Michael L. Leventhal, gynecologists in year 1935, the name of the condition is a misnomer as all the PCOS patients don't have polycystic ovaries [2]. The condition is characterized by hormonal imbalance, i.e. heightened androgen levels and metabolism problems. PCOS can result in the absence of ovulation, i.e. an ovulation because of hormonal imbalance resulting in irregular periods, enlarged ovaries with micro cysts and infertility [3]. In the majority of patients (75-85%) having PCOD there is clinically evident menstrual dysfunction can result in abnormal uterine bleeding [4]. The absence of ovulation can cause changes in levels of progesterone, estrogen, FSH and LH.PCOD are featured by increased LH, may cause muted FSH, high prolactin levels and increased gonadotropin–releasing hormone (GnRH) levels that can lead to increased free androgens in the body of patients [5] [6]. PCOS is mostly diagnosed on the basis of clinical symptoms, though ultrasonographic evidence of multiple micro cysts in the ovaries may help in diagnosis [7]. Diagnostic criteria may also include evaluation of morphological changes like the volume of the ovaries and antral follicular count as PCOS patients tends to have multiple follicles(>10) of 2–9 mm size and enlarged ovaries with volume of > 10 cm3[8][9][10] .In most of the patients (almost 70 %) the common symptoms of PCOS include:-

1. Acne
2. Hirustism
3. Baldness
4. Skin pigmentation
5. Obesity
6. Irregular periods
7. Infertility.

Almost 60-70% women suffering from PCOS have hirustism, 60-80% have elevated androgen levels and 70-80% have metabolic disorders like obesity and increased BMI [6] [11] [12]. Although the causes of PCOS are unknown, it is believed to be genetic in nature [13]. Almost 50-70% of the patients have significant insulin resistance in the body and can contribute to long term health problems like prediabetes, sleep apnea, greater risk of myocardial infarction (MI), dyslipidemia, high blood pressure, anxiety, depression, and endometrial cancer [14][15][16][17][18]. Insulin resistance or hyperinsulinemia also contributes to most symptoms of the disease and weight loss has been associated with regulated menstrual cycle in PCOS.

Treatment includes management of the disease and symptoms associated with it such as hirustism, acne, hormonal imbalance, infertility and obesity [19]. Lifestyle modifications, weight loss, suitable diet intake, regular exercise can help in the management of the disease resulting in reduced free androgen index and decreased biochemical hyperandrogenism[20-25]. It has been observed that symptoms become less severe with age and as women approach menopause. Medications used to address individual concerns like infertility, menstrual irregularities; increased hair growth etc may be used for birth control pills, anti-androgen medications, metformin, progesterone pills etc [26-29]. Although research studies were conducted to diagnose PCOS using different machine learning algorithms, but there is scope for improvement in accuracy and precision on the basis of clinical data [30-32].

2. DIFFERENT MACHINE LEARNING ALGORITHMS
There are a number of 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 under:

2.1 Logistic Regression
It is actually not a regression algorithm, but a classification used to estimate binary values. The given set of independent variable(s) decides the binary values and the probability of

occurrence of an event are predicted by fitting data to a logit function. The prediction of probability is performed and the outcome values lie between 0 and 1.

2.2 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.

2.3 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.

2.4 Classification and Regression Trees (CART)

One of the important types of algorithms is decision trees, which are mainly used for predictive modeling machine learning. It finds its application commonly in data mining with the soul purpose to predict the values of dependent variables (target) based on the values of several independent variables (input).

2.5 Random Forest

For a group of decision trees, the Random Forests are a trade mark term. The collection of decision trees forms a forest. Each tree votes for the class based on attributes.

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

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

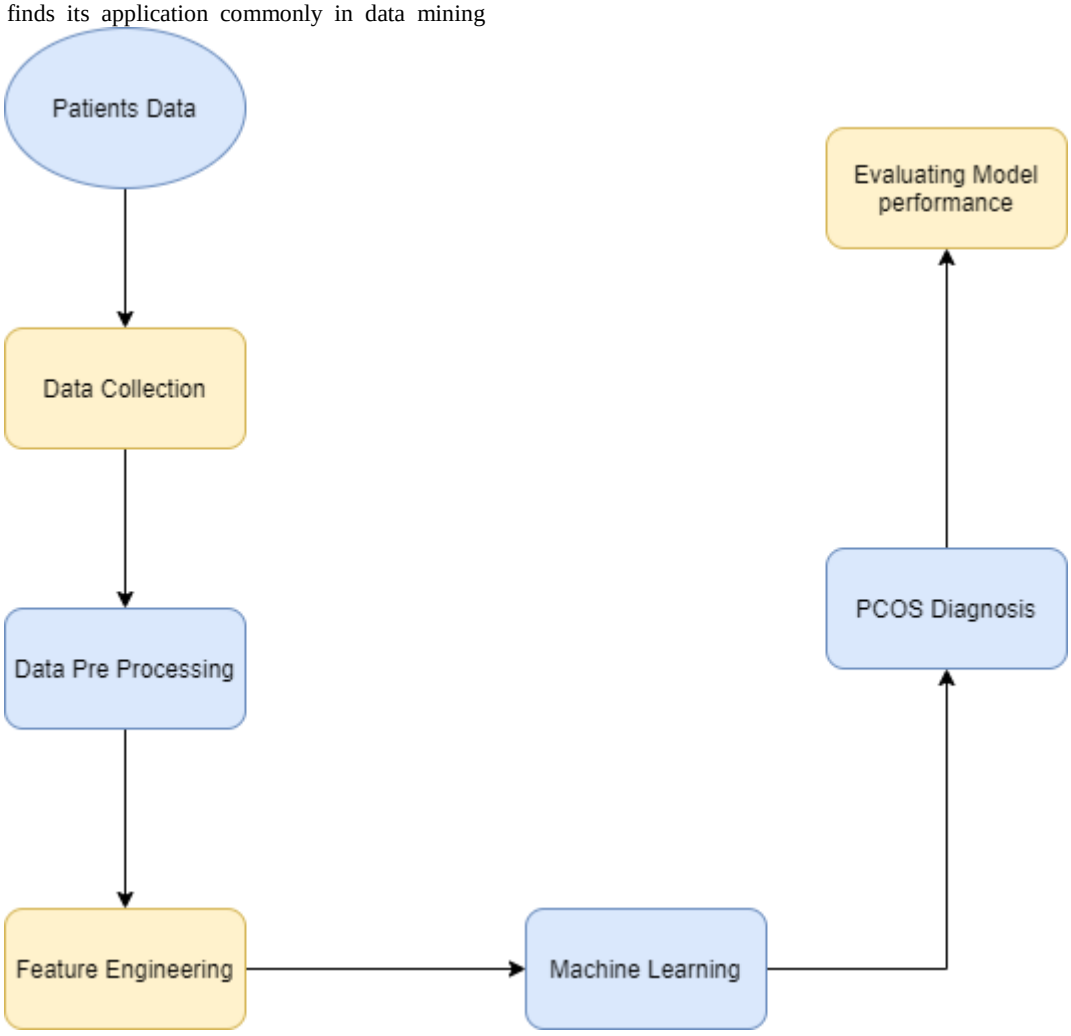

Fig 1: Methodology for the diagnosis of PCOS

Data collection: The data are collected from 10 different hospitals across Kerala, India and is available on Kaggle site.

Preprocessing of data: The data is preprocessed to find missing values and then used for diagnosing PCOS using different machine learning algorithms.

Classification: We are using R-language for implementation of classification algorithms, diagnosis and validation of model performance. R libraries used for the purpose were e1071, CARET, naivebayes, rpart, randomForest, klaR, ggplot2.

Five machine learning algorithms were trained and evaluated on random samples of data with 42 independent variables as symptoms to diagnose PCOS. The dependent variable PCOS strongly correlate with these independent variables and can

take two values 'Yes' or 'No'. The algorithms used include logistic regression, Random Forest, SVM (Support Vector Machine), CART (Classification and Regression Trees) and Naïve Bayes algorithm. The list of variables used for diagnosis of PCOS is depicted in Table 1.

Table 1: List of Variables used in diagnosis of PCOS

Dependent Variable	Value
PCOS	0-No,1-Yes
Independent Variables	Value
Age	Can take a valid numeric value
Weight	Can take valid numeric value
Height	Can take valid numeric value
BMI	Can take valid numeric value
Blood Group	Can take valid numeric value
Pulse rate(bpm)	Can take valid numeric value
RR (breaths/min)	Can take valid numeric value
Hb(g/dl)	Can take valid numeric value
Cycle(R/I)	Can take valid numeric value
Cycle length(days)	Can take valid numeric value
Marriage Status (Yrs)	Can take valid numeric value
Pregnant(Y/N)	Can take valid numeric value
No. of abortions	Can take valid numeric value
beta-HCG(mIU/mL)	Can take valid numeric value
beta-HCG(mIU/mL)	Can take valid numeric value
FSH(mIU/mL)	Can take valid numeric value
LH(mIU/mL)	Can take valid numeric value
FSH/LH	Can take valid numeric value
Hip(inch)	Can take valid numeric value
Waist(inch)	Can take valid numeric value
Waist:Hip Ratio	Can take valid numeric value
TSH (mIU/L)	Can take valid numeric value
AMH(ng/mL)	Can take valid numeric value
PRL(ng/mL)	Can take valid numeric value
Vit D3 (ng/mL)	Can take valid numeric value
PRG(ng/mL)	Can take valid numeric value
RBS(mg/dl)	Can take valid numeric value
Weight gain(Y/N)	Can take valid numeric value
hair growth(Y/N)	0-No,1-Yes
Skin darkening (Y/N)	0-No,1-Yes
Hair loss(Y/N)	0-No,1-Yes
Pimples(Y/N)	0-No,1-Yes
Fast food (Y/N)	0-No,1-Yes
Reg. Exercise(Y/N)	0-No,1-Yes
BP _Systolic (mmHg)	Can take valid numeric value
BP _Diastolic (mmHg)	Can take valid numeric value
Follicle No. (L)	Can take valid numeric value
Follicle No. (R)	Can take valid numeric value
Average F size (R) (mm)	Can take valid numeric value
Average F size (L) (mm)	Can take valid numeric value
Endometrium (mm)	Can take valid numeric value

4.1 Validation of models

The common performance evaluation metrics for validation of models include:

Accuracy: It is the proportion of the total number of predictions that were correct and can be calculated from the following equation:

$$\text{Accuracy} = \frac{Tx+Ty}{Tx+Fx+Ty+Fy}$$

Where, Tx= True Positives
Ty= True Negatives
Fx= False Positives
Fy= False Negatives [24][25][26].

Recall: is defined as the percentage of total relevant results correctly classified by the algorithm.
Recall=Tx/Tx+Fx
Precsion: refers to the percentage of the results which are relevant.
Precision=Tx/Tx+Ty
F-statistics:-is a metric that combines precision and recall and is calculated as the harmonic mean of precision and recall.
F_n = (1+n^2)*precision*recall/((n^2 * precision)+recall)

5. RESULTS

The logistic regression coefficients used to describe the relationship between the predictor and the response variables are shown in the table 2. The visualization of Residuals for Logistic Regression is shown in the figure 2. The ROC Curve of Logistic Regression is depicted in figure 3. The sensitivity is plotted in a function of specificity for different cut of points by Receiver Operating Characteristic (ROC) curve. The sensitivity/specificity pair with respect to specific decision threshold is indicated point wise in the ROC cure with 92% . The detailed comparison of the different classification algorithms employed in diagnosis the PCOS is shown in the figure 4. The comparison of algorithms using R-box and whisker plots are shown in figure 5. The comparison of Logistic Regression and SVM algorithms is shown in figure 6.

The comparison of RF (Random Forest) and SVM algorithms is depicted in figure 7. The comparison of Machine Learning Algorithms Using Parallel Plots is shown in figure 8. Figure 9 gives the comparative analysis of Machine Learning Algorithms with respect to four important features analyzed critically, which are "Recall, Precision, F1 and Accuracy" and all these four features are the best with Random Forest. Therefore, we come up with this conclusion that Random Forest is the best among the five algorithms used for the said study.

Table 2. Logistic Regression Coefficients

Concordance	Discordance	Kappa
0.92	0.07	110714.6

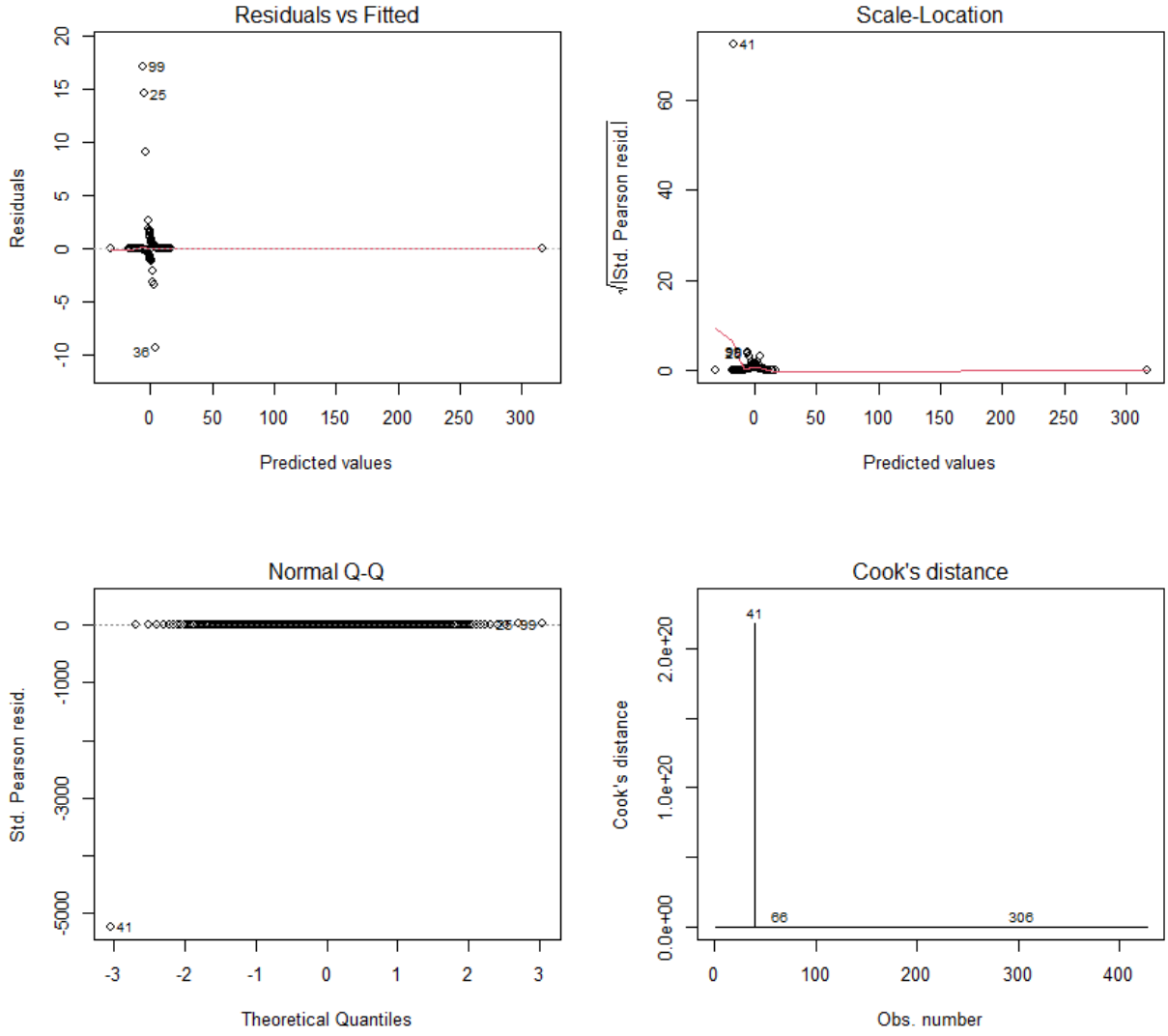

Fig 2: Visualization of Residuals for Logistic Regression

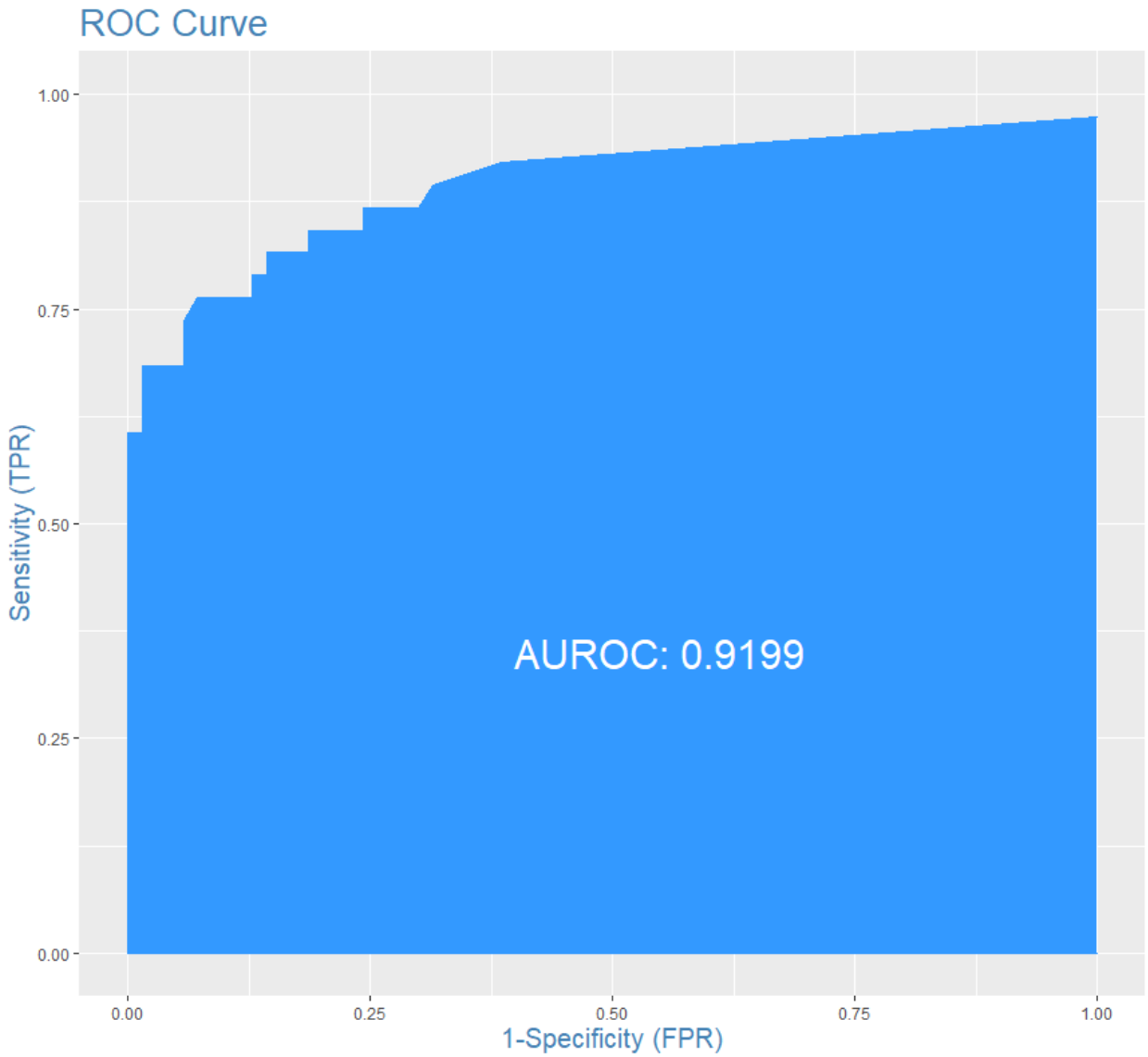

Fig 3: ROC Curve of Logistic Regression

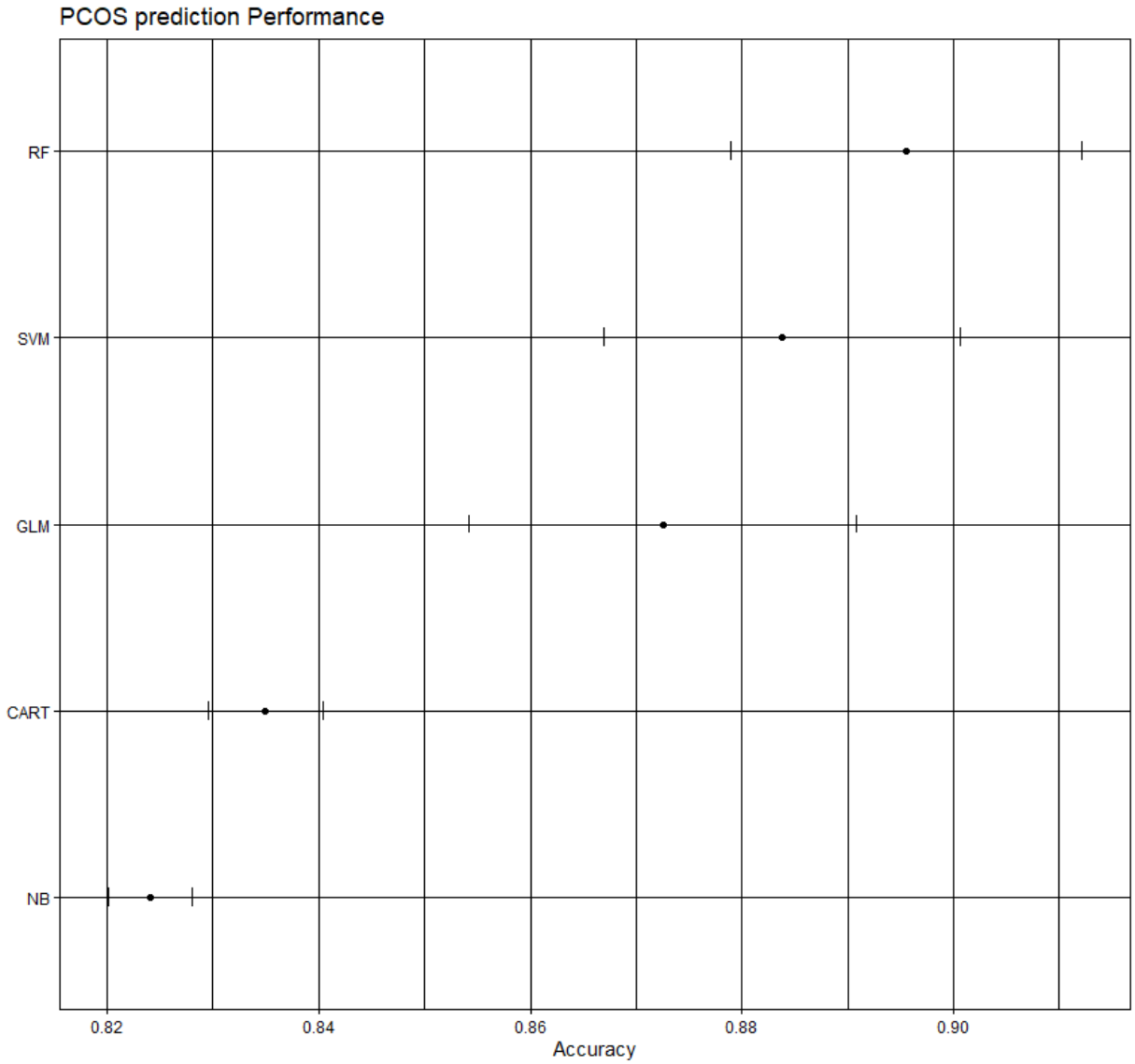

Fig 4: Comparison of different machine learning algorithms in diagnosis of PCOS

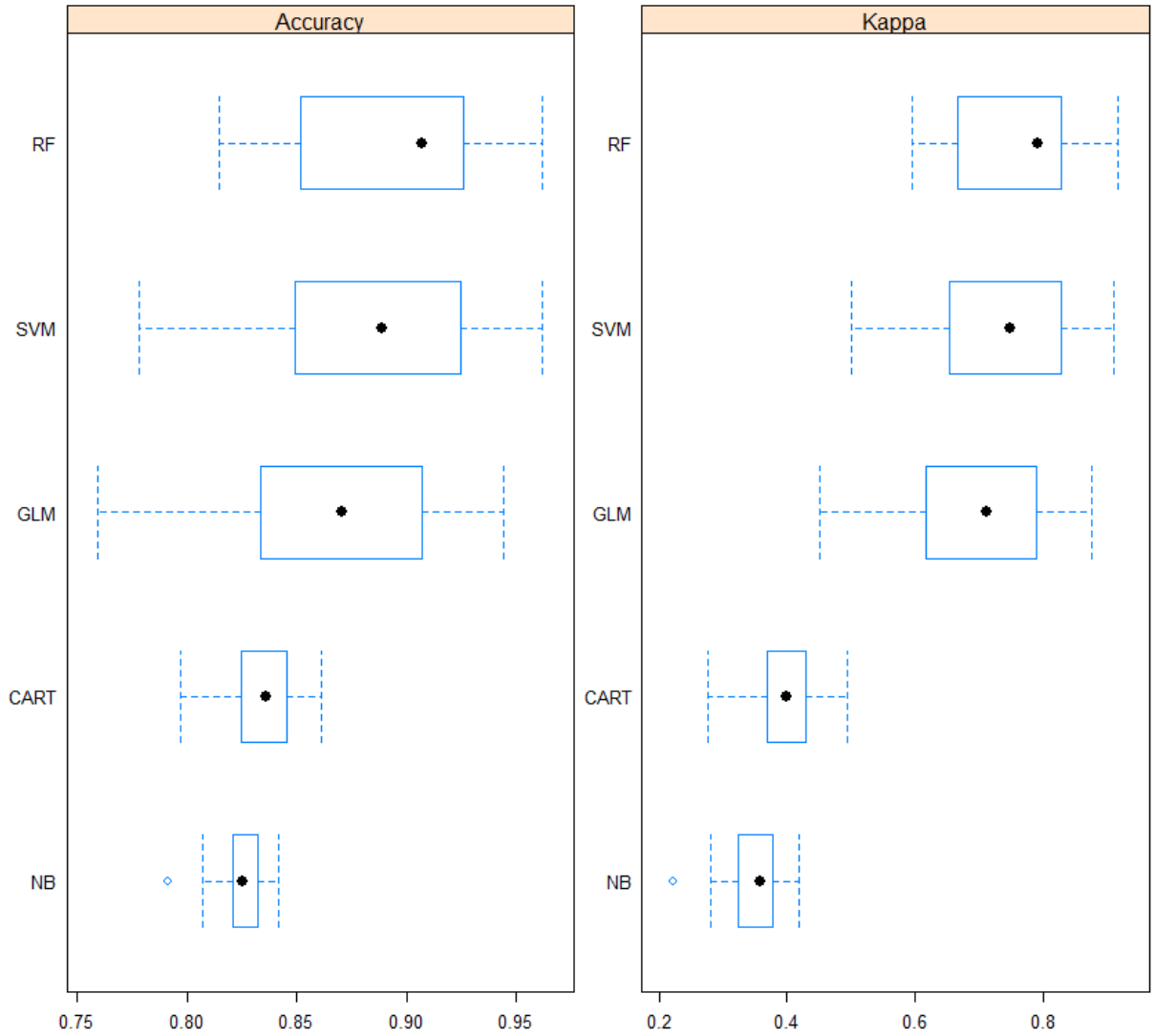

Fig 5: Comparison of machine learning algorithms in diagnosis of PCOS using R-box and whisker plots

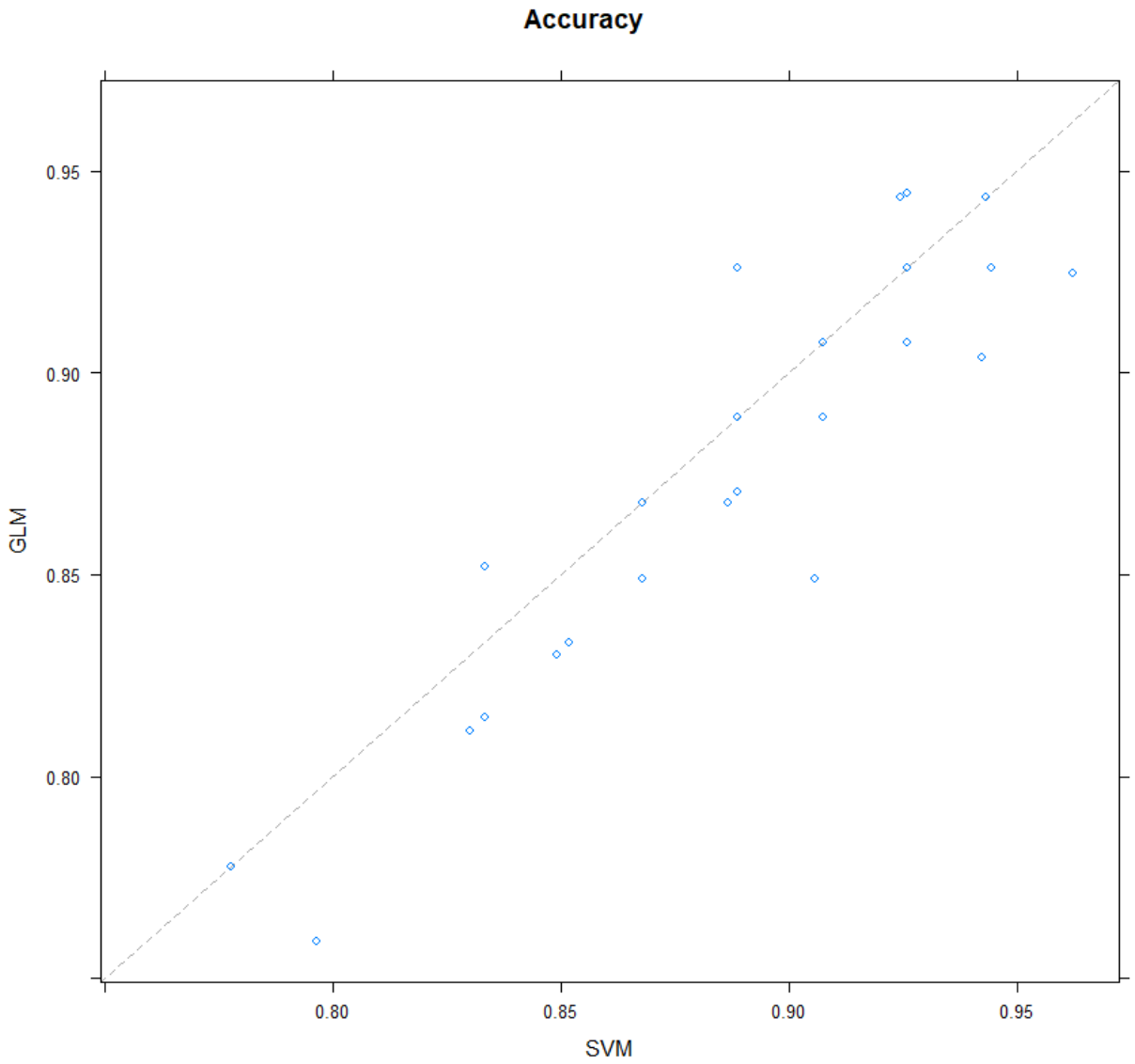

Fig 6: Comparison of Logistic Regression and SVM algorithms in diagnosis of PCOS

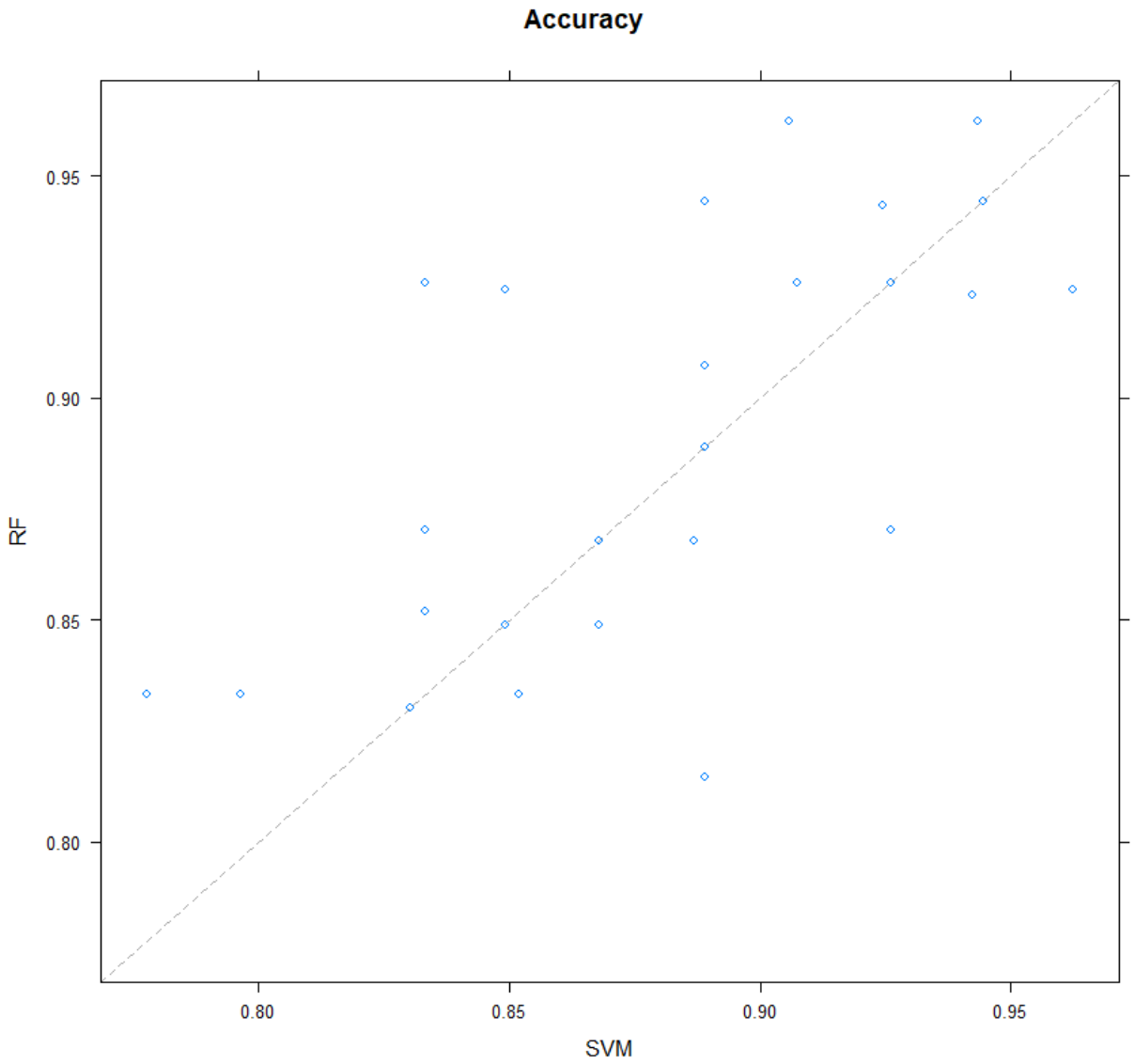

Fig 7: Comparison of RF (Random Forest) and SVM algorithms in diagnosis of PCOS

Fig 8: Comparison of accuracy of machine learning algorithms in PCOS diagnosis using Parallel Plots

Table 3. Comparative Analysis of Machine Learning Algorithms in diagnosis of PCOS

Algorithm	Logistic Regression	Support Vector Machine	Naïve Bayes	CART(Classification and Regression Trees)	Random Forest
Recall	0.91	0.95	0.76	0.94	0.95
Precision	0.98	0.95	0.94	0.92	0.96
F1	0.94	0.95	0.84	0.93	0.96
Accuracy	0.92	0.94	0.81	0.90	0.96

Table 3 displays the parameters for validation of machine learning models used for diagnosis of PCOS. In this study, we used four performance measures to indicate goodness of fit of the model, namely Recall, precision, F1 and accuracy. The values of recall,precision,F1 and accuracy indicate best performance of the Random Forest model on the given data.

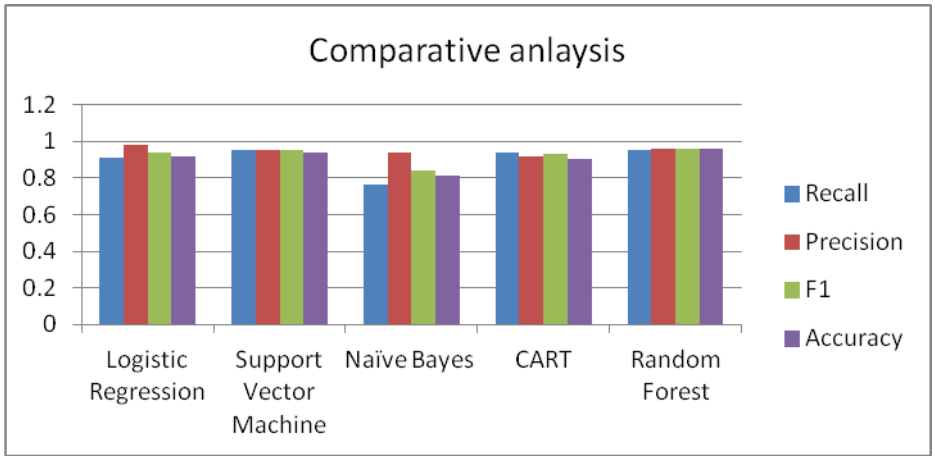

Fig 9: Comparative analysis of machine learning algorithms in diagnosis of PCOS

6. CONCLUSION

PCOS is a disorder caused by hormonal imbalance in the body of young women and is a very common problem affecting a substantial portion of women worldwide. Early diagnosis of the condition can help in the treatment and management of the disorder. We have used popular machine learning algorithms, i.e. SVM, CART, Naive Bayes Classification, Logistic Regression and Random Forest on the clinical data of patients diagnosed PCOS based on symptoms associated with the disease. The performance validation metrics recall, accuracy, precision, and F- statistics indicated best performance of the Random Forest algorithm in diagnosis of PCOS with an accuracy of 96% followed by SVM with accuracy of 95%.Thus, it is concluded that the Random Forest algorithm is the best suitable algorithm for diagnosis of PCOS on the given data. The future scope of the study can include use of different or large data sets for diagnosis of the disease.

7. REFERENCES

[1] G. N. Allahbadia and R. Merchant, "Polycystic ovary syndrome and impact on health," Middle East Fertil. Soc. J., vol. 16, no. 1, pp. 19–37, 2011, doi: 10.1016/j.mefs.2010.10.002.

[2] G. P. Rédei, "Polycystic Ovarian Disease (Stein-Leventhal syndrome)," in Encyclopedia of Genetics, Genomics, Proteomics and Informatics, Springer Netherlands, 2008, pp. 1528–1528.

[3] I. F. Stein and M. L. Leventhal, "Amenorrhea associated with bilateral polycystic ovaries," Am. J. Obstet. Gynecol., vol. 29, no. 2, pp. 181–191, 1935, doi: 10.1016/s0002-9378(15)30642-6.

[4] R. Rebar, Evaluation of Amenorrhea, Anovulation, and Abnormal Bleeding. MDText.com, Inc., 2000.

[5] S. M. Sirmans and K. A. Pate, "Epidemiology, diagnosis, and management of polycystic ovary syndrome," Clin. Epidemiol., vol. 6, no. 1, pp. 1–13, 2013, doi: 10.2147/clep.s37559.

[6] R. Azziz et al., The Androgen Excess and PCOS Society criteria for the polycystic ovary syndrome: the complete task force report, vol. 91, no. 2. 2009.

[7] W. U. Atiomo, S. Pearson, S. Shaw, A. Prentice, and P. Dubbins, "Ultrasound criteria in the diagnosis of polycystic ovary syndrome (PCOS)," Ultrasound Med. Biol., vol. 26, no. 6, pp. 977–980, Jul. 2000, doi: 10.1016/S0301-5629(00)00219-2.

[8] A. H. Balen et al., "Andrology: Polycystic ovary syndrome: The spectrum of the disorder in 1741 patients," Hum. Reprod., vol. 10, no. 8, pp. 2107–2111, 1995, doi: 10.1093/oxfordjournals.humrep.a136243.

[9] B. Y. Jarrett et al., "Impact of right–left differences in ovarian morphology on the ultrasound diagnosis of polycystic ovary syndrome," Fertil. Steril., vol. 112, no. 5, pp. 939–946, Nov. 2019, doi: 10.1016/j.fertnstert.2019.06.016.

[10] T. Williams, "Diagnosis and Treatment of Polycystic Ovary Syndrome," Jul. 2016. Accessed: Jul. 25, 2020. [Online]. Available: http://familydoctor.org/familydoctor/en/diseases-conditions/polycystic-ovary-syndrome.html.

[11] D. A. Ehrmann et al., "Prevalence and predictors of the metabolic syndrome in women with polycystic ovary syndrome," J. Clin. Endocrinol. Metab., vol. 91, no. 1, pp. 48–53, Jan. 2006, doi: 10.1210/jc.2005-1329.

[12] S. Sam, "Obesity and polycystic ovary syndrome," Obesity Management, vol. 3, no. 2. pp. 69–73, Apr. 2007, doi: 10.1089/obe.2007.0019.

[13] M. P. Lauritsen, P. F. Svendsen, and A. N. Andersen, "Diagnostic criteria for polycystic ovary syndrome," Ugeskr. Laeger, vol. 181, no. 15, pp. 671–679, 2019, doi: 10.1016/S1701-2163(16)32915-2.Diagnostic.

[14] M. J. Himelein and S. S. Thatcher, "Depression and body image among women with polycystic ovary syndrome," J. Health Psychol., vol. 11, no. 4, pp. 613–625, Jul. 2006, doi: 10.1177/1359105306065021.

[15] A. Dunaif, K. R. Segal, W. Futterweit, and A. Dobrjansky, "Profound peripheral insulin resistance, independent of obesity, in polycystic ovary syndrome," Diabetes, vol. 38, no. 9, pp. 1165–1174, 1989, doi: 10.2337/diab.38.9.1165.

[16] R. S. Legro, "Polycystic Ovary Syndrome and Cardiovascular Disease: A Premature Association?," Endocr. Rev., vol. 24, no. 3, pp. 302–312, Jun. 2003, doi: 10.1210/er.2003-0004.

[17] L. M. Liao, J. Nesic, P. M. Chadwick, K. Brooke-Wavell, and G. M. Prelevic, "Exercise and body image distress in overweight and obese women with polycystic ovary syndrome: A pilot investigation," Gynecol. Endocrinol., vol. 24, no. 10, pp. 555–561, 2008, doi: 10.1080/09513590802288226.

[18] R. A. Wild, "Dyslipidemia in PCOS," in Steroids, Mar. 2012, vol. 77, no. 4, pp. 295–299, doi: 10.1016/j.steroids.2011.12.002.

[19] E. C. Costa et al., "Aerobic Training Improves Quality of Life in Women with Polycystic Ovary Syndrome," Med. Sci. Sports Exerc., vol. 50, no. 7, pp. 1357–1366, Jul. 2018, doi: 10.1249/MSS.0000000000001579.

[20] M. A. Karimzadeh and M. Javedani, "An assessment of lifestyle modification versus medical treatment with clomiphene citrate, metformin, and clomiphene citrate-metformin in patients with polycystic ovary syndrome," Fertil. Steril., vol. 94, no. 1, pp. 216–220, Jun. 2010, doi: 10.1016/j.fertnstert.2009.02.078.

[21] I. Almenning, A. Rieber-Mohn, K. M. Lundgren, T. S. Løvvik, K. K. Garnæs, and T. Moholdt, "Effects of high intensity interval training and strength training on metabolic, cardiovascular and hormonal outcomes in women with polycystic ovary syndrome: A pilot study," PLoS One, vol. 10, no. 9, Sep. 2015, doi: 10.1371/journal.pone.0138793.

[22] D. R. Chizen et al., "The 'pulse' diet & PCOS," Fertil. Steril., vol. 102, no. 3, p. e267, Sep. 2014, doi: 10.1016/j.fertnstert.2014.07.908.

[23] D. S. Kiddy et al., "Improvement in endocrine and ovarian function during dietary treatment of obese women with polycystic ovary syndrome," Clin. Endocrinol. (Oxf)., vol. 36, no. 1, pp. 105–111, 1992, doi: 10.1111/j.1365-2265.1992.tb02909.x.

[24] H. H. Mehrabani, S. Salehpour, B. J. Meyer, and F. Tahbaz, "Beneficial effects of a high-protein, low-glycemic-load hypocaloric diet in overweight and obese women with polycystic ovary syndrome: A randomized controlled intervention study," J. Am. Coll. Nutr., vol. 31, no. 2, pp. 117–125, Apr. 2012, doi: 10.1080/07315724.2012.10720017.

[25] F. Giallauria et al., "Exercise training improves autonomic function and inflammatory pattern in women with polycystic ovary syndrome (PCOS)," Clin. Endocrinol. (Oxf)., vol. 69, no. 5, pp. 792–798, Nov. 2008, doi: 10.1111/j.1365-2265.2008.03305.x.

[26] U. A. Ndefo, A. Eaton, and M. R. Green, "Polycystic ovary syndrome: A review of treatment options with a focus on pharmacological approaches," P T, vol. 38, no. 6, pp. 336–355, Jun. 2013, Accessed: Jul. 25, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3737989/.

[27] L. Radosh, "Drug Treatments for Polycystic Ovary Syndrome," Apr. 2009. Accessed: Jul. 25, 2020. [Online]. Available: www.aafp.org/afp.

[28] G. Ladson et al., "The effects of metformin with lifestyle therapy in polycystic ovary syndrome: A randomized double-blind study," Fertil. Steril., vol. 95, no. 3, Mar. 2011, doi: 10.1016/j.fertnstert.2010.12.002.

[29] A. Gambineri et al., "Treatment with flutamide, metformin, and their combination added to a hypocaloric diet in overweight-obese women with polycystic ovary syndrome: A randomized, 12-month, placebo-controlled study," J. Clin. Endocrinol. Metab., vol. 91, no. 10, pp. 3970–3980, 2006, doi: 10.1210/jc.2005-2250.

[30] V. Deepika, "Applications of Artificial Intelligence Techniques in Polycystic ovarian syndrome Diagnosis," J. Adv. Res. Technol. Manag. Sci., vol. 1, no. 3, pp. 59–63, 2019, doi: 10.1007/978-981-13-2640-0.

[31] A. Saravanan and S. Sathiamoorthy, "Detection of Polycystic Ovarian Syndrome : A Literature Survey," vol. 7, no. 2, pp. 46–51, 2018.

[32] V. Krishnaveni, "A Roadmap to a Clinical Prediction Model With Computational Intelligence for PCOS," Vol. Ix, No. Ii, Pp. 177–185.

8. AUTHOR'S PROFILE

Malik Mubasher Hassan is working as Principal University Polytechnic Baba Ghulam Shah Badshah University (BGSBU) Rajouri (J&K), India. He had served as Head of the Department ITE BGSBU for about ten years. He received M.Tech degree from NIT Srinagar in 2007 and is pursuing Ph.D form the same institute. His research interests include communication systems, optical wireless, computer Networks, data mining and ICT. He has published number of research articles/papers in the reputed national/international journals.

Tabasum Mirza: She has received a Masters Degree in Computer Applications from University of Kashmir in 2008. She is presently working as Lecturer in the Department of Computer Science, School Education, Government of Jammu and Kashmir, India. She has a 6.5 years experience of working in JK Bank Pvt. Ltd. Her specialization is software Engineering, Java Programming, Data Mining and e-learning. She has published various research articles/papers in the reputed national/international journals.