International Journal of Advanced Research in Engineering and Technology (IJARET)
Volume 11, Issue 11, November 2020, pp. 1872-1881, Article ID: IJARET_11_11_174
Available online at http://iaeme.com/Home/issue/IJARET?Volume=11&Issue=11
ISSN Print: 0976-6480 and ISSN Online: 0976-6499
DOI: 10.34218/IJARET.11.11.2020.174

© **IAEME** Publication **Scopus** Indexed

PREDICTION OF POLYCYSTIC OVARIAN SYNDROME WITH CLINICAL DATASET USING A NOVEL HYBRID DATA MINING CLASSIFICATION TECHNIQUE

Neetha Thomas
Research Scholar, Kongunadu Arts and Science College, Tamilnadu, India

Dr. A. Kavitha
Associate Professor, Kongunadu Arts and Science College, Tamilnadu, India

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is a female hormone disorder which sorely affects women health by developing symptoms like irregular menses, infertility, obesity, hyperandrogenism, alopecia, may lead to another severe health risk like metabolic syndrome, type 2 diabetes mellitus, Cardio Vascular Diseases etc. It affects 5-10% of women in their puberty. The objective of the study is to predict the occurrence of PCOS before getting worse. Data mining holds excellent prospective to ameliorate health care; several studies have already undergone to foreshadow the danger of PCOS in women. The proposed system is a novel hybrid structure to determine the chances of PCOS that coalesce navies Bayes and artificial neural network algorithm to produce the best result.

Keywords: PCOS, Machine learning, Navies Bayes, Artificial neural network, Hybrid structure.

Cite this Article: Neetha Thomas and A. Kavitha, Prediction of Polycystic Ovarian Syndrome with Clinical Dataset Using a Novel Hybrid Data Mining Classification Technique, *International Journal of Advanced Research in Engineering and Technology,* 11(11), 2020, pp. 1872-1881.
http://iaeme.com/Home/issue/IJARET?Volume=11&Issue=11

1. INTRODUCTION

Data mining is the inspection and study of extensive data to discover meaningful information. It's recognized under the area of data science, which is a well-known field of study, provides tremendous productivity in all areas, one such is Medical mining, a wide area of research, encompasses different kinds of mining methods to solve problems for diagnostics and treatment, and also understanding the progression of the disease with more personalized and effective care. It transforms clinical dataset to knowledgeable information which assists in clinical decision making and personalized medicine. Different techniques are used in data mining for prediction and decision making for different types of diseases, including kidney disease, diabetes, infertility, cancer, heart disease, etc. The classifiers should be carefully chosen based on the selected problem and dataset available. In this paper, we are focusing on the prediction of Poly Cystic Ovarian Syndrome (PCOS).

Polycystic ovary syndrome (PCOS) is a complex condition that is most often diagnosed by the presence of any of the two from the following criteria: hyperandrogenism, ovulatory dysfunction, and polycystic ovaries [1]. This is an ill-understood mystifying condition with no definite cure. Young adolescent girls experience the full range of symptoms from irregular menses, amenorrhea, menorrhagia, hirsutism, acne, skin pigmentation, alopecia, anxiety, depression, ovarian cysts [2]. All through the sign of PCOS varies from person to person but most common symptoms like infertility, hirsutism, mensural disorders remain same so that physicians can clinically diagnose PCOS if the patient has anovulation, oligomenorrhea, and patient have hyperandrogenism (e.g., hirsutism, acne).then their ovaries may be polycystic, [3]. The lack of ovulation changes the levels of estrogen, progesterone, FSH, and LH. Thus androgen, the male hormones, which should present a little in women's body increases, i.e., hyperandrogenism, and cause to derange menstrual cycle, it leads to getting fewer menstruation and ovulation for women with PCOS than usual [4].To know the severity of the disease regarding the size and measure of follicles in ovaries and to verify the thickness of uterus lining patients should definitely undergo ultrasound scanning or further diagnostic steps. The ovaries are 1½ to 3 times larger than normal ovaries.

We use the popular Analytics and Data Mining Software tool MATLAB R2018a. The organization of the paper follows. In Sect II, we describe the materials we collected and methods we used In Sect III Data mining methodologies and performance evaluation parameters are explained Sect. IV, we explained the proposed structure in Sect. V, we present our conclusions.

2. MATERIALS AND METHODS

2.1. Data Source

The study was administered based on the clinical diagnosis and prescription by doctors from various clinics and hospitals in and around the Thodupzha Municipality. Based on the research studies, it is proven that PCOS diagnosis can be primarily confirmed with clinical investigations, and later on to understand the severity of the disease, the patient should undergo pelvic ultrasound measures. Based on this an e-form was pre-designed to input the entries for both married and unmarried women. In order to find the best model to predict the chances of occurring polycystic ovarian syndrome in women population, a real dataset has been used. From the 208 women considered for the study 114 having PCOS and rest were examined as usual.

The following set of attributes is used for predicting the Polycystic Ovarian Syndrome for the women at an earlier stage.

Table 1 PCOS Dataset

Attributes	Values
Painful menstruation (Dysmenorrhoea)	YES/NO/SOMETIMES
Acne	YES/NO
Severe pain before a week of the period	YES/NO/SOMETIMES
Absence of menstruation for more than two months (Amenorrhoea)	YES/NO/SOMETIMES
Severe vomiting and pain during period	YES/NO/MAYBE
Excess hair growth on arms, face, chest, abdomen, back or any other visible parts on the body (Hirsute)	YES/NO/A LITTLE
Obesity	YES/NO
Infertility	YES/NO
Light menstrual periods	YES/NO/SOMETIMES
Prolonged bleeding (Menorrhagia)	YES/NO/SOMETIMES
Thinning of hair on the head	YES/NO
Sleep problems	YES/NO/SOMETIMES
Mood swings, AnxietyAnxiety	YES/NO/SOMETIMES

In this paper, we have meticulously described the predictive analysis and showcase the probability for chances of PCOS disease. The attributes values for PCOS disease prediction is prepared purely based on clinical investigations proposed by doctors; The database consists of 13 main attributes and 208 instances, which can be classified as data, string and numbers, the candidate can enter 'yes', 'no', 'maybe', 'sometimes', 'A little' according to their symptoms shown.

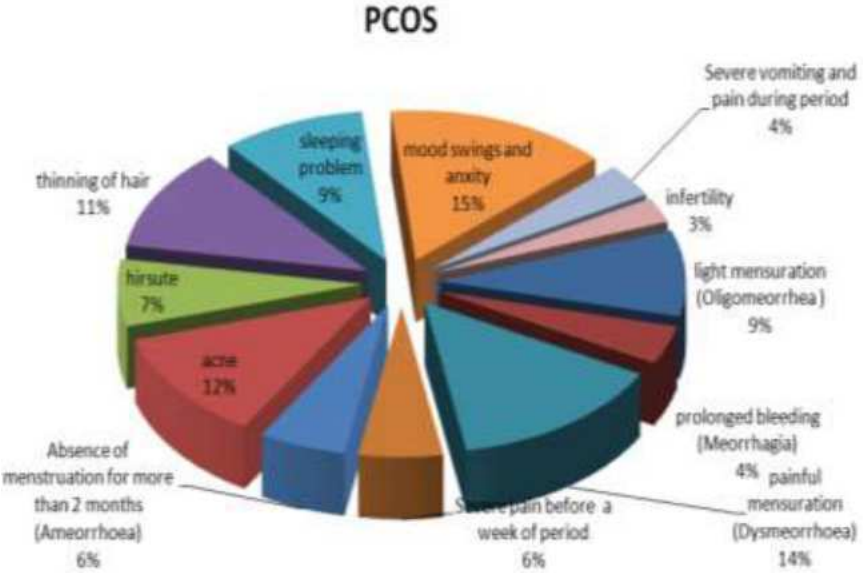

Figure 1 Data Attributes with the percentage of occurrence

The entries are converted into values ranging from 0-4 for the ease of assessing attribute values. Then the complete data set is categorized into three groups according to the severity of the disease. Fig 1 depicts clinical dataset with symptoms and the percentage of occurrences pictorially. There are four binary attributes and eight categorical attributes, as shown above in Table1.

2.2. Data partitioning procedure

Training Data	Testing Data
70%	30%

From the 208 instances of data, 70% of PCOS data instances in the data set are assigned for the training set, and the remaining 30% are selected for testing purpose. Partitioning a data set into training and testing, it typically related to the model, which includes different kinds of patterns from the training data and evaluating that model accuracy using highly similar but different instances from the test data [5]. In our work, we chose 8:3 ratio as it provided satisfactory results.

3. HYBRID CLASSIFICATION FRAMEWORK AND PERFORMANCE PARAMETERS

Several research studies are already undergone in data mining methodologies to predict the chances of PCOS in women. Palak Mehrotra et al. [6] conducted a survey of PCOS; they concluded that Bayesian classifier gives higher accuracy than the logistic regression. Rethinavalli et al. [7] proposed the NFRS and Hybrid Classification Algorithm for the prediction of PCOS. B. Vikas et al. [8], had done a study on PCOS with FIM, Apriori algorithm. Later they revised reviews [9], by coined that Navies Bayes gave better performance. In this paper, we explained about the methodologies Navies Bayes, Artificial Neural Network and a hybrid technique

3.1. Naive Bayesian

Naive Bayesian or Bayes' Rule is a classification approach representing independence among predictors through theorization, which is the basis for many machine-learning and data mining methods[10]. It has the capability to create models making predictions, also provides a ceaseless way for exploring and understanding data. Navies Bayes classifier has a pivotal role in medical data mining. The performances are very high as the attribute values are not dependent on one other[11]. In Naive Bayes Algorithm:

$$P\ (c|v) = (P\ (v|c)\ P(c)/P\ (v))$$

Assume v represent the parameters/features in the dataset. i.e. v=(v1, v2…,vn), predicting n measurements on tuple from n attributes, the elements, can be mapped to as Painful menstruation, Acne, Obesity, etc. After substituting for **v** and expanding we get. P is the probability, variable **c** is the class variable(predict, PCOS), which represents to predict the disease on given the conditions.

$$P\ (c|v_1, v2, v3 \ldots vn) = \frac{P(v_1|c)\ P(v_2|c)\ P(v_3|c)\ P(v_4|c) \ldots\ldots\ldots\ P\ (v_n|c)\ P(c)}{P\ (v_1)\ P\ (v_2)\ P\ (v_3)\ P\ (v_4) \ldots\ldots\ P\ (v_n)}$$

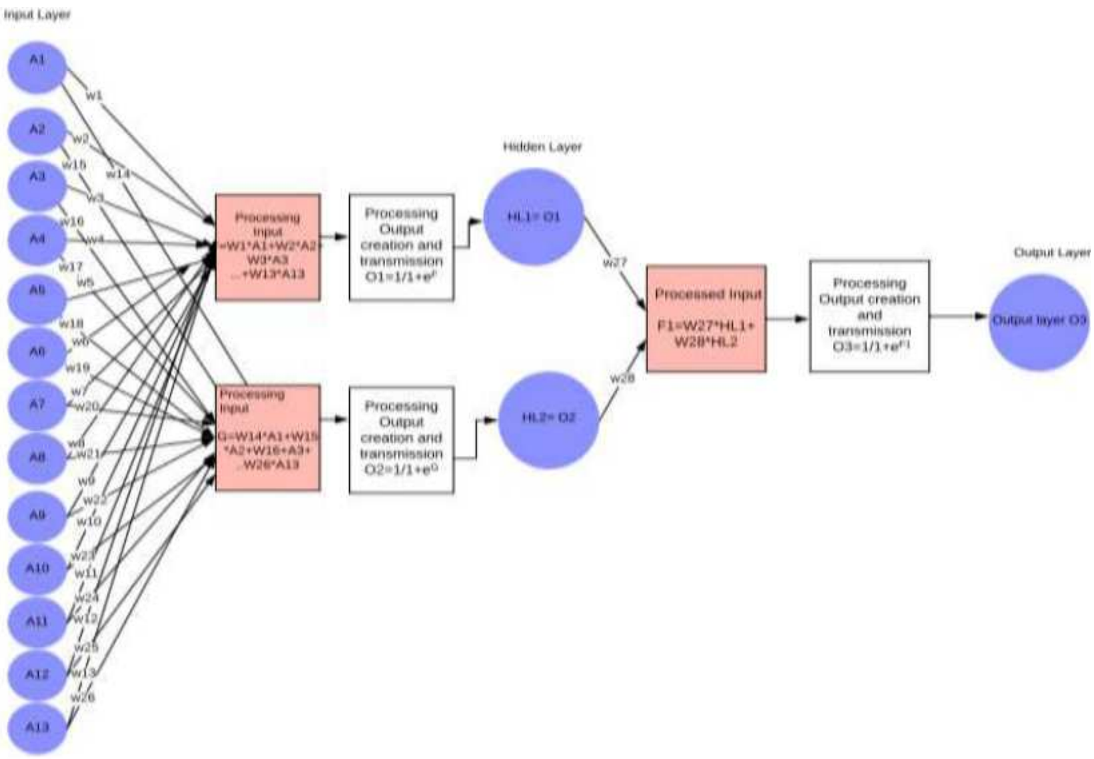
Input Layer
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
w1
w2
w3
w4
w5
w6
w7
w8
w9
w10
w11
w12
w13
w14
w15
w16
w17
w18
w19
w20
w21
w22
w23
w24
w25
w26
Processing Input =W1*A1+W2*A2+ W3*A3 ...+W13*A13
Processing Output creation and transmission O1=1/1+e^F
Processing Input G=W14*A1+W15 *A2+W16+A3+ ..W26*A13
Processing Output creation and transmission O2=1/1+e^G
Hidden Layer
HL1= O1
HL2= O2
w27
w28
Processed Input F1=W27*HL1+ W28*HL2
Processing Output creation and transmission O3=1/1+e^F1
Output Layer
Output layer O3

Figure 2 Neural Network Simulation of PCOS Dataset

Add the parameter values from the dataset and reserve them into the equation. Therefore, we can reform the proportionality as:

$$P\ (c/v_1,\ldots\ldots v_n) \propto P(c)\ \prod_{i=1}^{n} P(v_i/c)$$

Using the MATHLAB function for constructing Naive Bayes classifier fitcnb is used [12]. Fitcnb (h, v) returns a multiclass naive Bayes model, trained by predictors h and data values v and predict, enumerate the probabilities of a categorical class variable given with independent predictor variables using Bayes rule. x= predict (NBmodel, TF) predicts the output of an identified model, ahead using input-output data history from TF.

3.2. Artificial Neural Network

The Neural Network Simulator performs backpropagation and generates the testing output for ANN classifier. The output is calculated as a function [13].

Using the MATHLAB function for constructing ANN classifier, transit (), Asymmetric sigmoid transfer function, which is the activation function between two hidden layers, consists of 10 input nodes each. The above diagram shows an architectural representation of ANN for PCOS prediction fig2, here the 13 attributes are marked as input attributes. Hidden node distils significant patterns from the inputs and transfers to the next layer, generates outputs O1, O2 for the final prediction of output O3.

Using the sigmoid function O3 =1/(1+exp(-F1)), Where O3 is output neuron, and F1 is the product of weights of .output from the first hidden layer and second hidden layer. The activation function sigmoid creates a non-linear relation with all input attributes and generates functional output.

Feed-forward backpropagation network object, newff consist of L1 layers using the dot prod (w1=dot (i1, i2)) weight function, then perform net sum which is a net input function and specified transfer functions. The first layer has weights coming from the input.

Each subsequent layer has a weight coming from the previous layer. The last layer is the network output. Each layer executes, according to Nguyen-Widrow, initialization algorithm that boots a layer's weights and biases. Acclimation is done with sequential order incremental training function that updates values with the specified learning function.

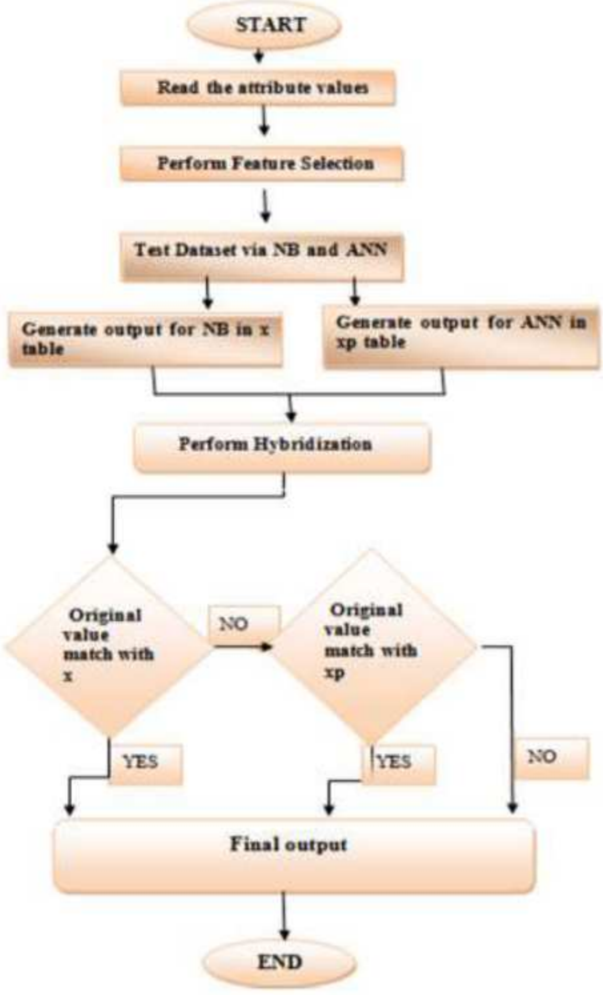
START
Read the attribute values
Perform Feature Selection
Test Dataset via NB and ANN
Generate output for NB in x table
Generate output for ANN in xp table
Perform Hybridization
Original value match with x
NO
Original value match with xp
YES
YES
NO
Final output
END

Figure 3 Hybrid Classification Approach

3.3. Hybrid classification framework

The concept of Hybridization come up by incorporating the two classifiers by analyzing their predicted values; each value is compared with the original table and generated the best hybrid classification. Navies Bayes algorithm has the advantage of requiring less information regarding systems variables, predicted the output with least accuracy for this dataset, and it is far below our expectation, while the probability of values generated by ANN classifier was better which is further our notion. Hybrid classification algorithm combining Bayesian classifier and Neural Network classifier has cross analyzed both the classifier predicted values and produced the superior out turn. The execution part of the proposed machine learning technique is mentioned in the flowchart fig 3.

3.4. Performance Evaluation Metrics

For statistically measuring the performance of the PCOS classification evaluation, Performance parameters like Accuracy, Precision, Recall, F-measure and specificity are calculated for all the classifiers, which are generally coupled with a Binary classification kind of test. True positives (TP), True negatives (TN), False positive (FP), False Negative (FN) are calculated for every test dataset for evaluating the classifiers.

Specificity depict the negative tuples which are calculated as;

$$\frac{TN\ frequency}{FP\ frequency + TN\ frequency}$$

Precision depicts the ratio of TP with all possible positives results.

$$\frac{TP\ frequency}{TP\ frequency + FP\ frequency}$$

Accuracy shows appropriate classification percentage of tuples.

$$\frac{TP\ frequency + TN\ frequency}{TP\ frequency + FN\ frequency + TN\ frequency + FP\ frequency}$$

Recall depicts the ground truth of reality check in tuples.

$$\frac{TP\ frequency}{TP\ frequency + FN\ frequency}$$

F-measure Provide a realistic measure of a test's performance with respect to precision and recall.

$$\frac{Precesion * Recall}{Precesion + Recall}$$

4. RESULT AND DISCUSIONS

The study of 208 cases was collected from various hospitals and scanning centre at Thodupuzha. The dataset consist of women of reproductive age, among them, 94 were average 26 instances with the likelihood to PCOS, the remaining 114 cases reported as PCOS. Here we listed a graphical representation of results with respect to the dataset.

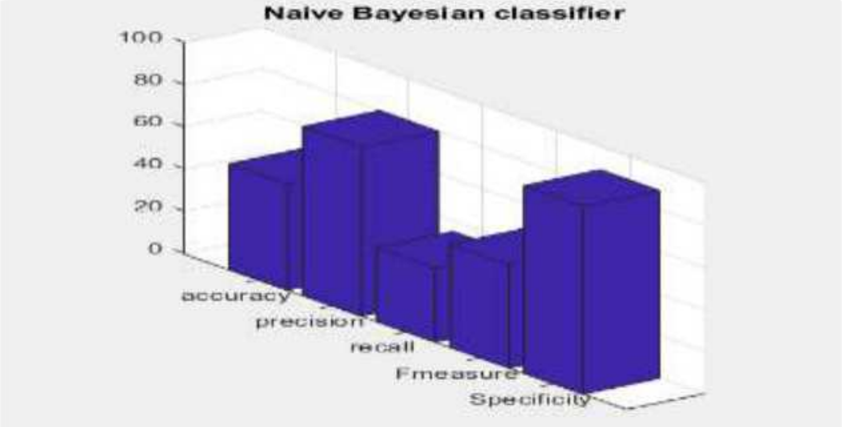

Figure 4 Performance analysis of NB Classifier

Table 2 Comparison of Classifies with Performance Metrics

Classifiers	Accuracy	Precision	Recall	F-measure	Specificity
Navies Bayes	0.50	0.80	0.3529	0.4897	0.8928
Artificial Neural Network	0.7596	1.0	0.6764	0.8070	1.0
Hybrid Classifier	0.9534	1.0	0.9375	0.9677	1.0

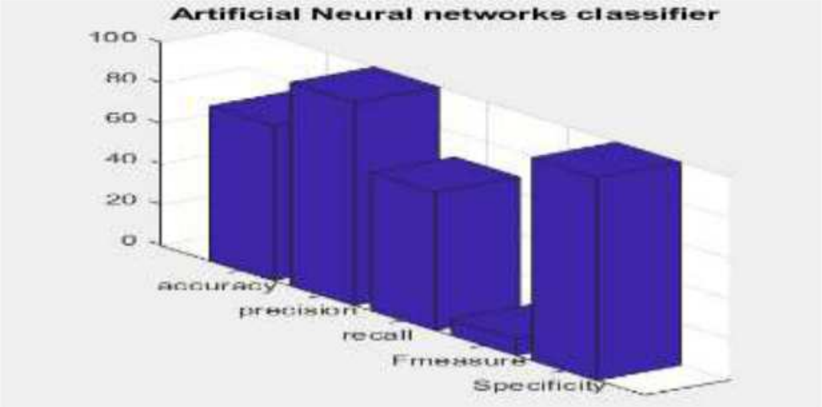

Figure 5 Performance analysis of ANN Classifier

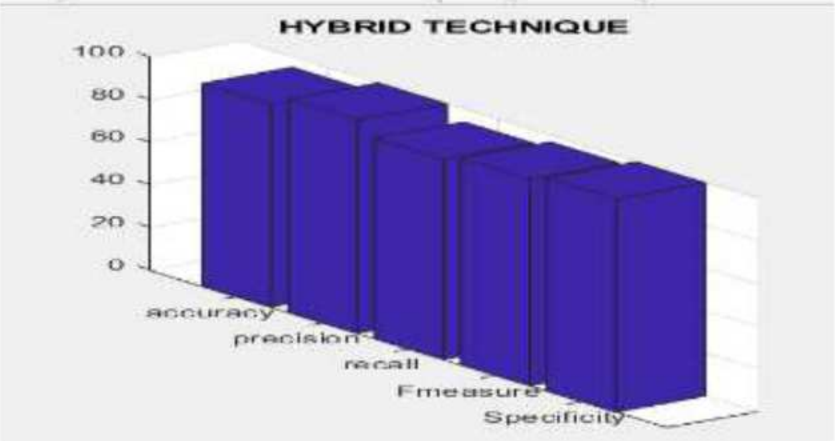

Figure 6 Performance analysis of Hybrid Classifier

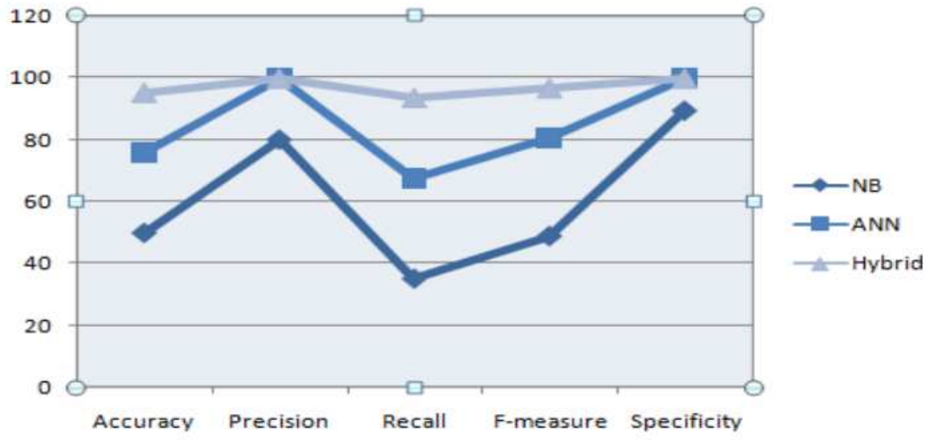

Figure 7 Graphical Representation of Classifiers

5. CONCLUSION

PCOS is becoming a seriously ill condition in women population these days; it affects 5-10% of women in their childbearing age. Proper medication, along with lifestyle management, can reduce the seriousness of syndrome. We have introduced a new machine learning computational method with respect to clinical characteristics which aid doctors for examining the patients, thereby saving time and can diagnose PCOS in the beginning stage itself. In the study, we have incorporated both Navies Bayes and Artificial Neural Network classifiers to attain the desired upshot. The Navies Bayes algorithm gives reasonable accuracy, considering PCOS dataset, followed by Artificial Neural Network Algorithm with the sub-optimal accuracy of 75 per cent following the Artificial Neural Network is the Hybrid Classifier with an optimal accuracy of 95 percentage Eventually, from these algorithms we can elucidate that the Hybridization of classifiers produced the best result when comparing to others.

The future scope is wide open, Transvaginal sonography (TVS) has a significant role in the treatment and diagnosis of PCOS. The polycystic pattern which is depicted by the presence of many cysts ranging from 2-9 mm in diameter should identify and can undergo image processing techniques for further understanding*; Also it provides detailed results regarding the* severity of disease on affected patient.

ACKNOWLEDGEMENT

We would like to pay our heartfelt gratitude to Dr Meena soman, Dr Elizebeth joy, Chazikattu Hospital, Department of Obst. & Gyn., Thodupuzha, Dr Ajaykumar, MD, Chief-Reproductive Medicine, Archana Hospital, Dr Krishnadas MD Consultant Radiologist and all staff of radiology department, Medical Diagnostic Center (MDC) for generosity in accumulating the data and apprising us about the relationships among various symptoms. Any Errors found our own mistake please should not tarnish the reputation of the *esteemed* person

REFERENCES

[1] Tracy Williams MD, Rami Mortada MD, Samuel Porter MD, “Diagnosis and Treatment of Polycystic Ovary Syndrome”, Am Fam Physician. 2016 Jul 15; 94(2):106-113

[2] Choudhary A, Jain S, Chaudhari P. Prevalence and symptomatology of polycystic ovarian syndrome in Indian women: is there a rising incidence?. Int J Reprod Contracept Obstet Gynecol 2017; 6:4971-5

[3] LEE RADOSH, MD, "Drug Treatments polycystic Ovary Syndrome", The Reading Hospital and Medical Center, Reading, Pennsylvania Am Fam Physician. 2009 Apr 15; 79(8):671-676.

[4] Neetha Thomas, Dr A.Kavitha," A literature inspection on polycystic ovarian morphology in women using data mining methodologies", International Journal of Advanced Research in Computer Science, Volume 9, No. 1, January-February 2018.

[5] Liu, H., Cocea, M. Semi-random partitioning of data into training and test sets in a granular computing context. Granul. Comput. 2, 357–386 (2017). https://doi.org/10.1007/s41066-017-0049-2

[6] Palak Mehrotra, Jyotirmoy Chatterjee, Chandan Chakraborty School of Medical Science and Technology Indian Institute of Technology

[7] S. Rethinavalli and M. Manimekalai, “A Novel Hybrid Framework for Risk Severity of Polycystic Ovarian Syndrome”, I J C T A, 9(27), 2016, pp. 19-27,© International Science Press

[8] B. Vikas(&), B. S. Anuhya, K. Santosh Bhargav, Sipra Sarangi, and Manaswini Chilla" Application of the Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome (PCOS), "Springer Nature Singapore Pte Ltd. 2018

[9] B. Vikas(&), B. S. Anuhya, K. Santosh Bhargav, Sipra Sarangi, and Manaswini Chilla "Critical Study of Polycystic Ovarian Syndrome (PCOS) Classification Techniques", IJCEM International Journal of Computational Engineering & Management, Vol. 21Issue4, and July 2018 ISSN (Online): 2230-7893

[10] Palak Mehrotra, Jyotirmoy Chatterjee, Chandan Chakraborty, BiswanathGhoshdastidar, Sudarshan Ghoshdastidar G D Institute for Fertility Research, "Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques",_Published in 2011 Annual IEEE India Conference

[11] Monika Gandhi, Dr Shailendra Narayan Singh," Predictions in Heart Disease Using Techniques of Data Mining", 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management(ABLAZE-2015)

[12] Computational Statistics Handbook with MATLAB, By Wendy L. Martinez, Angel R. Martinez pg- 365

[13] Ankita Dewan, Meghna Sharma," Prediction of Heart Disease Using a Hybrid The technique in Data Mining Classification", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)