IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022

Prediction of PCOS Using Ensemble Learning Algorithms

[1*] B. V. Ramana, [2] T. Ravi Kumar, [3] B. R. Sarath Kumar
[1] Dept of IT, Aditya Institute of Technology and Management, Tekkali, AP, India.
[2] Dept of CSE, Aditya Institute of Technology and Management, Tekkali, AP, India.
[3] Dept of CSE, Lenora College of Engineering, Rampachodavaram, A.P, India
[*]Corresponding Author: ramana.bendi@gmail.com

Abstract

Polycystic Ovary Syndrome (PCOS) is endocrine disorder effecting many women's their in reproductive age system. This may result infertility and an ovulation, and it also causes hormonal imbalance which leads to a delayed or even absent menstrual cycle. Approximately 5-10 % of reproductive age (15-49 years) women are suffered from this problem. Women who are being suffered from this disease are being effected from following symptoms weight gain, facial hair growth, acne, hair loss, skin darkening and irregular periods. Now a days we have so many methodologies and treatments to predict other diseases in earlier stage, but when it comes to PCOS the existing treatments are insufficient to predict this disease in earlier stage. PCOS due to overlapping the follicles, inheritance of the equipment lack of operator knowledge as it is large experiment dependent procedure To deal with this problem infertility, diabetes mellitus, cardiovascular diseases, our proposed system which can help early detection and prediction of PCOS treatment from an optimal and minimal set of parameters. In this paper we are going to work with some machine learning algorithms those are Ad boost, Cat boost, XGboost, Gboost and Bagging. All these techniques are tested with applying Recursive Feature Elimination (RFE) methodology.

Keywords: Polycystic Ovary Syndrome, Machine Learning, XGboost, Ad boost, Cat boost, Gboost and Bagging.

1. Introduction

Polycystic Ovarian Syndrome also called as Stein–Eventual syndrome is an endocrine disorder affecting 5 to 10 percent of women in reproductive age (12-45 years). The disease was first discovered by Irving F.Stein,Sr., and Michael L.Leventhal, gynecologists in year1935,the name of the condition is a misnomer as all the PCOS patients don't have polycystic ovaries. The condition is characterized by hormonal imbalance, i.e. heightened androgen levels and metabolism problems. PCOS can result in the absence of ovulation,

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022**

i.e. an ovulation because of hormonal imbalance resulting in irregular periods, enlarged ovaries with micro cysts and infertility. In the majority of patients (75-85%) having PCOD there is clinically evident menstrual dysfunction can result in abnormal uterine bleeding. 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. PCOS is mostly diagnosed on the basis of clinical symptoms, though ultrasonographic evidence of multiple micro cysts in the ovaries may help in diagnosis. 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 cm

1.1. Impacts of PCOS:

Irregular Periods: A lack of ovulation prevents the uterine lining from shedding every month.Some women with PCOS get fewer than eight periods a year or none at all. Infrequent, irregularor prolonged menstrual cycles are the most common sign of PCOS. Although some women with PCOS how is regular periods high level of androgens and too much insulin in their bodie scan disrupt the monthly cycle of ovulation and menstruation of many women with PCOS.

Heavy bleeding: The uterine learning builds up for a longer period of time. So, the periods you do get can be heavier than normal PCOS can also cause heavy fast-flowing periods, and sometimes they come with large blood clot Abnormal hair growth or hair loss: Excessive growth of hair on the face, chest, back, or buttock sis called hirsutism. This condition is the effect of high production of the male hormone androgen, which is what results in unwanted facial and body hair. On the flipside, excess androgen scan also result in hair loss or acne.

Weight gains: Up to 80% of women with PCOS are over weighted or how obesity. Male pattern baldness. Hair on the scalp gets thinner and may fall out. Darkening of the skin. Dark patches of skin can form in body creases like those on the neck, In the groin, and under the breasts. Losing weight can be a healthy way to keep your cholesterol and blood sugar levels in check, both of which are important if you have PCOS.

Infertility: Women with PCOS have a hormonal imbalance and metabolism problems that may affect their overall health and appearance. Because ovulation does not occur regularly,

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022

periods become irregular and increased levels of hormones such as testosterone can affect egg quality, inhibit ovulation, lead to insulin resistance, and increase the risk disorders such as gestational diabetes.

Darkening of the skin: Dark patches of skin can form on body creases like those on the neck, in the groin, and under the breasts. These PCOS discolored skin patches are often found in the fold so of kin and commonly occurred around Armpits, Groin, neck, Vulva, elbows, knees, knuckles, lips.

Headaches: Hormone changes can trigger headaches in some women. The surging hormones that cause PCOS can give you headaches, too.

Technology is changing every outlook of our lives making remarkable transformations in the health care industry. Now-a-days technology and humans are working hand in hand. For example, robots performing surgeries once seemed a fiction but now they are performing critical and complex surgeries in hospitals.

1.2. Adaboost Algorithm

AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1split. These trees are also called Decision Stumps.

The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage. It’s no wonder then that CERN recognized it as the best approach to classify signals from the Large Hadron Collider. This particular challenge posed by CERN required a solution that would be scalable to process data being generated at the rate of 3 peta bytes per year and effectively distinguish an extremely rare signal from background noises in a complex physical process. XGBoost emerged as the most useful, straight forward and robust solution.

1.3. Gboost Algorithm

Gradien Boosting is a machine learning technique used in regression and classification tasks,among others. It gives a prediction model in the form of an ensemble of weak prediction models,which are typically decision trees. When a decision tree is a weak learner, the resulting algorithm is called gradient-Boosted trees; it usually outperforms random forest. A gradient-Boosted trees model is built in a stage-wise fashion as in other

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022

boosting methods, But it arbitrary differentiable loss function.

1.4. Catboost Algorithm

According to the CatBoost documentation, CatBoost supports numerical, categorical, and text features but has a good handling technique for categorical data.The CatBoost algorithm has quite a number of parameters to tune the features in the processing stage. Boosting in CatBoost refers to the gradient boosting machine learning. Gradient boosting is a machine learning algorithm for classification and regression problems. Which produces a prediction model in an ensemble of weak prediction models, typically decision trees.

1.5. Bagging Algorithm

Bagging also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual datapoints can be chosen more than once. After several data samples are generated, these weak models are then trained independently and depending on the type of task—regression or classification, for example—the average or majority of those predictions yield a more accurate estimate. As a note, the random forest algorithm is considered an extension of the bagging method, using both bagging and feature randomness to create an uncorrelated forest of decision trees.

2. Related Work

This section provides a brief literature work on PCOS. Table I summarizes the literature survey. In the literature of PCOS [1-2], follicles detection has been described in two methods, one is stereo logy 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 folli clean 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 back propagation 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)–Euclid an distance, and (3)neural network-

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022**

learning vector quantization (LVQ). In [8], follicles are segmented using region growing scheme. This technique tests whether the neighbor of initial seeds should beaded to segmentation region.

The segmentation process of ultrasound image of PCOS has been done in [9, 10] using edge detection. The authors of consider a median filter to remove noise from PCOS images. Different types of Machine Learning algorithms [used classifying the diseases 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 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 level on each side of the threshold.

In the Canny edge detection, a computational approach to Canny edge detection was presented in. 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.

3. Methodology

In this study, We considered XGboost as our proposed methodology. The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage. It's no wonder then that CERN recognized it as the best approach to classify signals from the Large Hadron Collider. This particular challenge posed by CERN required a solution that would be scalable to process data being generated at the rate of 3 peta bytes per year and effectively distinguish an extremely rare signal from background noises in a complex physical process. XGBoost emerged as the most useful, straightforward.

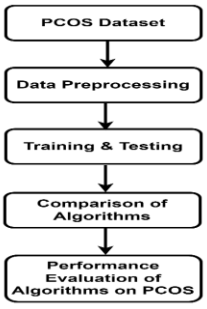


Fig 1: XGBoost Prediction

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022**

3.1. Dataset Overview

The PCOS prediction dataset was used to perform the study. There were 541 rows and 44 columns in his dataset. The value of the output column PCOS (Y/N) is either 1 or 0. The number 0 indicates that no PCOS risk was identified, while the value 1 indicates that a PCOS risk was detected. The probability of 0 in the output column (stroke) exceeds the possibility of 1 in the same column in this dataset. 354 rows alone in the PCOS (Y/N) column have the value 1, whereas 187 rows have the value 0.

3.2. Data Pre-processing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

3.3. Need of Data Pre-processing:

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022

4. Results

In this experimentation the accuracy of AdaBoost, XGBoost, G Boost, Bagging and CatBoost were calculated for both training and testing before applying Recursive Feature Elimination (RFE) and represented in table 1. The confusion matrices for are represented in Fig 2, Fig 3, Fig 4, Fig 5 and Fig 6 respectively. The comparison of all algorithms are represented in Fig 7.

Table 1: Accuracy table before applying RFE

S.No	Name of the Algorithm	Train Accuracy (%)	Precision	F1 score	Recall score	Test Accuracy (%)	Precision	F1 score	Recall score
1	AdaBoost	100	100	100	100	89.95	85	90	96
2	XGBoost	100	100	100	100	92.41	84	91	98
3	G Boost	99.80	99	99	100	90.41	86	90	96
4	Bagging	100	100	100	100	92.69	89	92	97
5	CatBoost	98.62	98	98	98	90.41	86	90	95

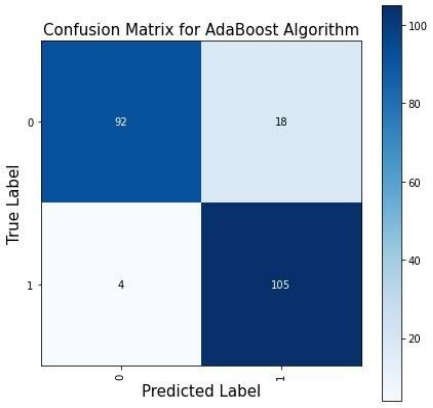


Fig 2: Confusion matrix for Adaboost

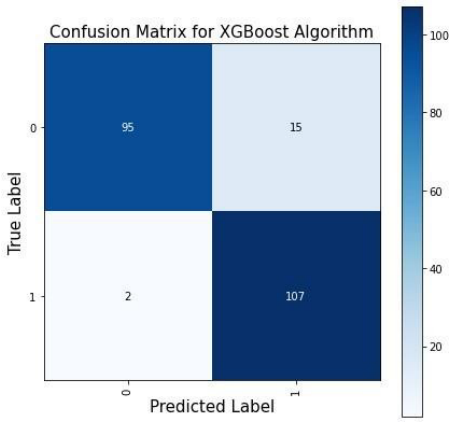
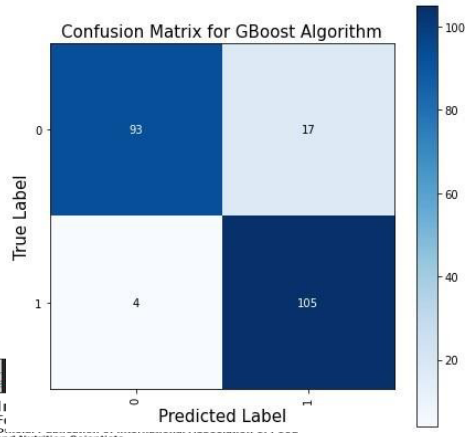
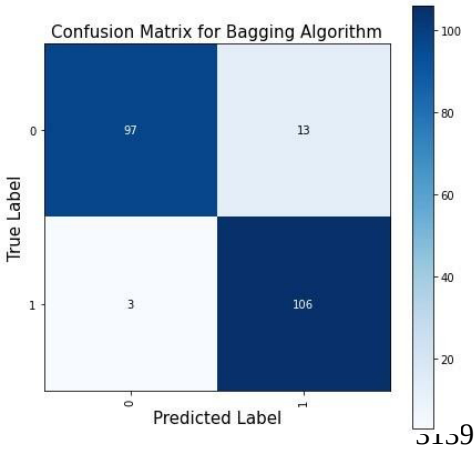


Fig 3: Confusion matrix for XGBoost





ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022

Fig 4: Confusion matrix for GBoost

Fig 5: Confusion matrix for Bagging

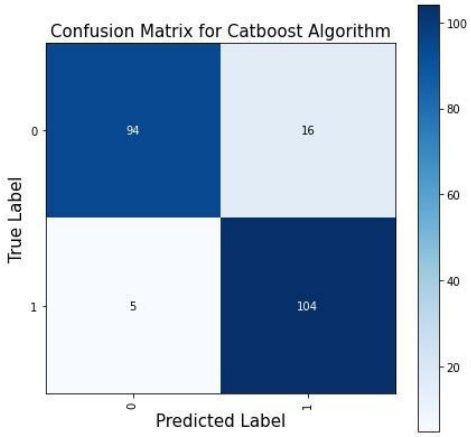


Fig 6: Confusion matrix for Catboost

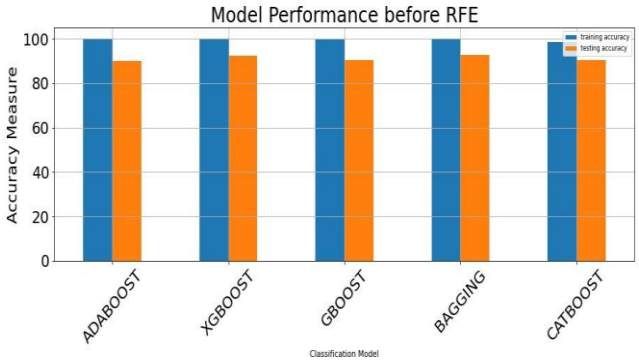
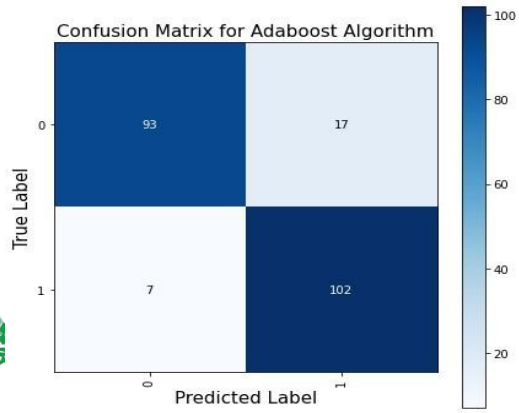


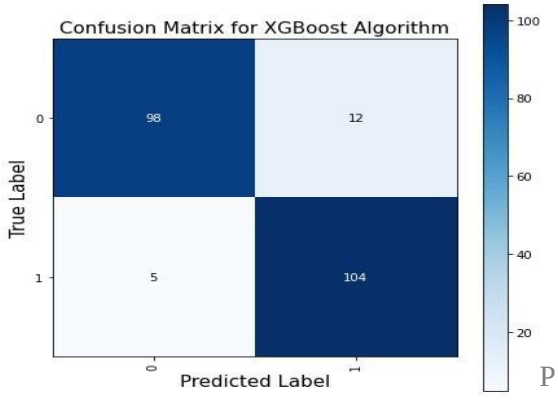
Fig 7: Comparison of accuracies before Applying RFE

Table 2: Accuracy table after applying RFE

S.No	Name of the Algorithm	Train Accuracy (%)	Precision	F1 score	Recall score	Test Accuracy (%)	Precision	F1 Score	Recall score
1	AdaBoost	100	100	100	100	89.04	85	89	93
2	XGBoost	100	100	100	100	93.15	89	92	95
3	G Boost	99.21	99	99	99	92.69	89	92	97
4	Bagging	100	100	100	100	90.41	86	90	95
5	CatBoost	98.82	98	98	98	91.78	87	92	97

The accuracy of AdaBoost, XGBoost, G Boost, Bagging and CatBoost were calculated for both training and testing after applying Recursive Feature Elimination (RFE) and represented in table 2. The confusion matrices for are represented in Fig 8, Fig 9, Fig 10, Fig 11 and Fig 12 respectively. The comparison of all algorithms are represented in Fig 13.





ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022**

Fig 8: Confusion matrix for AdaBoost

Fig 9: Confusion matrix for XGBoost

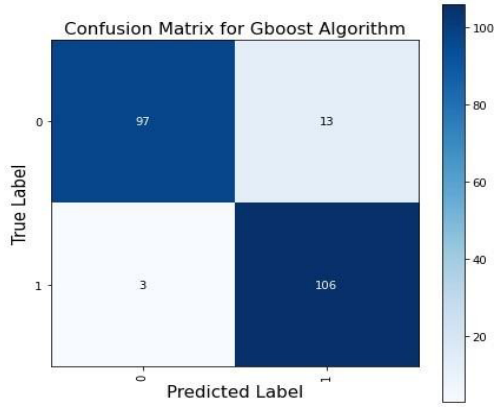


Fig 10: Confusion matrix for GBoost

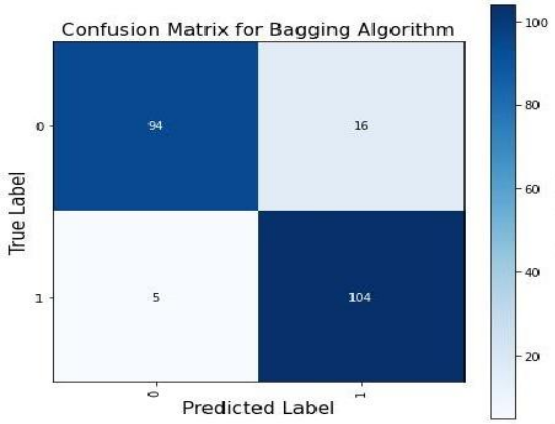


Fig 11: Confusion matrix for Bagging

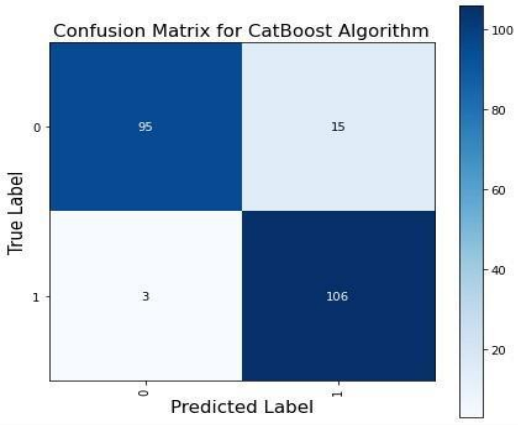


Fig 12: Confusion matrix for CatBoost

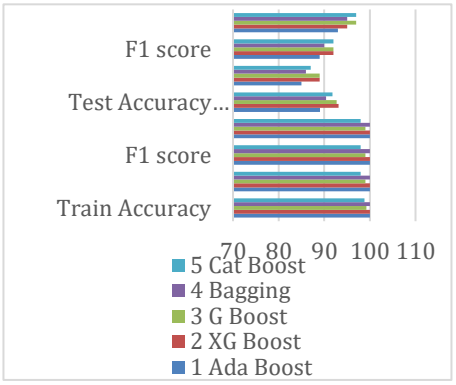


Fig 13: Comparison of accuracies after Applying RFE

5. Conclusion

So, Finally I conclude by saying that, this project PCOS prediction using machine learning is very much useful in every women's life and it is mainly more important for the health care sector, because they are the one that daily uses these systems to predict the PCOS of the patients based on their general information that they are been through. Now a day's health industry plays major role in predicting the PCOS of the patients so this is also some

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022**

kind of help for the health industry. If health industry adopts this project, then the work of the doctors can be reduced and they can easily predict the PCOS of the patient. Thus as we get maximum accuracy in XGboost, So we choose XGboost as our final algorithm to find possibilities of PCOS disease.

References

[1] R.Pasqualietal.,"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.13652265.2010.03956.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", 3rdInternational Conference on Science & Engineering in Mathematics, Chemistry and Physics2015(ScieTech 2015), vol. 622, 2015.

[4] U. N. Wisesty, J. Nasri and Adiwijaya, "Modified backpropagation algorithm for policy sticovary syndrome detection based on ultrasound images", Recent Advances on Soft Computing and Data Mining The Second International Conference on Soft Computing and Data Mining (SCDM2016), 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

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Oct 2022**

images",3rd International Conference on (IEEE) Information and Communication Technology (ICoICT), pp396-401,2015.

[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] M. Jayanthi Rao, R. Kiran Kumar: Follicle Detection in Digital Ultrasound Images using BEMD and Adaptive Clustering Algorithms, Innovative Product Design and Intelligent Manufacturing Systems, 2020, 651-659, SPRINGER. (SCOPUS), [ISBN:978-981-15-2695-4]

[10] M. Jayanthi Rao, R. Kiran Kumar: Method for Follicle Detection and Ovarian Classification using Geometrical Features, Journal of Advanced Research in Dynamical and Control Systems, 2019, 11(2), 1249-1258. (SCOPUS), [ISSN: 1943-023X]