American Journal of Electronics & Communication

Anisha Saha et. al., American Journal of Electronics & Communication, Vol. IV (2), 1-6

A Comparative Study to Predict Polycystic Ovarian Syndrome (PCOS) Based on Different Models of Machine Learning Technique

[1]Anisha Saha, [2]Aporna Roy, [3]Barsha Chakraborty, [4]Bidisha Saha, [5]Dipwanita Chowdhury
[anisha.sahaiembca2023, royaporna4, cbarsha119, bidishasaha50864, dipwanitachowdhury]@gmail.com
BCA 3rd Year, IEM, Kolkata

[6]Prof. Manab Kumar Das, [7]Prof. Soham Goswami
[manab.das, soham.goswami]@iem.edu.in
Assistant Professor, IEM, Kolkata

***Abstract* - Polycystic ovarian syndrome (PCOS) is a common endocrine disorder affecting women of reproductive age worldwide, characterized by excess production of androgens. This can result in ovarian abnormalities and a range of associated health risks, including infertility, heart issues, diabetes, and uterine cancer. However, the diagnosis of PCOS can be challenging due to the varied symptoms in different women and the time and cost involved in biochemical tests and ovarian scanning. To address this, researchers have proposed a method that predicts the likelihood of PCOS based on a minimal set of criteria, including weight, BMI, cycle length, and hormone levels. Using five machine learning algorithms, they tested the method on a dataset of 541 patients and found that the Random Forest and Support Vector Machine models had the highest accuracy in predicting PCOS. Such a system could aid in early detection and encourage individuals at risk to seek medical attention. Dataset is split into a 70/30 ratio, meaning that 70% of the dataset's data are used to train the model and 30% are used to test it. In this paper, we suggested a novel stack model with a 90% accuracy that is composed of four machine learning classifiers: Random Forest, Support Vector Machine, Naive Bayes, and Logistic Regression. Testing data accuracies for the models of Logistic Regression, Random Forest, Support Vector Machine, K Nearest Neighbor, Naive Bayes, Stack Model are 88%, 91%, 90%, 69%, 86% and 90% respectively. As a result, the models with the highest accuracy on the testing data are the Random Forest model and Stack Model.**

Keywords – Machine Learning Algorithms, PCOS, Ensemble Learning, Feature Selection, PRASOON KOTTARATHIL Dataset.

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a medical condition which causes hormonal disorder in women in their childbearing years. In the case of PCOS, ovaries can bulge and sometimes may have multiple small cyst formations (immature follicles). PCOS women have high levels of male hormones and insufficient female hormones, leading to alteration in their menstrual cycle or even absent menstrual cycle. Women with PCOS majorly suffer from excessive weight gain, facial hair growth, acne, hair loss, skin darkening and irregular periods leading to infertility.

The healthcare industry could undergo numerous revolutions because of artificial intelligence (AI). AI's capacity to analyze massive amounts of data correctly and rapidly is a key advantage. This enables medical workers to make better choices about patient care, such as individualized treatment plans and quicker evaluations. AI can also help with medical imagery by offering more precise and effective picture analysis, which lowers the chance of misdiagnosis. Additionally, real-time patient health monitoring by AI-powered gadgets enables early identification of possible health issues and prompt medical assistance. AI can also aid in the finding of new drugs by analyzing vast quantities of data and spotting potential treatments that may have gone unnoticed. Overall, AI has the potential to boost productivity, lower expenses, and enhance service quality in the healthcare industry.

The general factors of PCOS such as heredity, fast food, diet habits, involvement in physical exercise, BMI etc. The long-term effects of polycystic ovaries can cause significant ailments such endometrial hyperplasia, coronary disease, and type 2 diabetes mellitus. Studies have shown that it can also result in various malignancies including uterine or breast cancer in women who are fertile. Identifying PCOS is tricky due to all these manifestations, gynecological, clinical and metabolic parameters involved in diagnosing it. So, the time and financial expenses have become a hardship to the patients.

Our contributions in this paper-

- By integrating the Random Forest, Support Vector Machine, Naive Bayes, and Logistic Regression models, we have created a stack model that provides 97% accuracy on training datasets and 90% accuracy on testing datasets.
- Additionally, we compared the accuracies of Logistic Regression, Random Forest, Support Vector Machine, K Nearest Neighbor, and Naive Bayes model and the most accurate model was the random forest one.
- We used the Prasoon Kottarathil Dataset for our research and also preprocessed the data by managing null values and extracting features using methods like Pearson's correlation coefficient and the k-best

algorithm approach to improve the performance of our models.

II. Literature Survey

This study provides an in-depth analysis of PCOS and explores the use of image processing and machine learning techniques to aid in its diagnosis and potential automation. A range of analytical techniques have been utilized to detect and analyze PCOS.

The intention for conducting PCOS research is multifaceted, and can include addressing a major public health issue, advancing scientific knowledge, developing personalised care, and improving patient health outcomes. PCOS research can aid in the identification of risk factors, biomarkers, and other indicators that predict the development of PCOS, allowing for earlier interventions and personalised treatment plans. The ultimate goal is to improve the quality of life for women suffering from PCOS. To gain a comprehensive understanding of PCOS, it is necessary to reference established diagnostic criteria and standards.

M. Sumathi et al. [1] constructed a CNN image processing model for disease classification using ultrasound images. Feature extraction was performed using the watershed algorithm, and parameter measurement was carried out using OpenCV. The model achieved a high accuracy of 85% based on performance factors. Irfan Talib et al. [2] found that elevated insulin levels are a leading factor in the development of PCOS. Their study examined the diverse effects of insulin resistance in women with polycystic ovaries. R.M.Dewi et al. [3] proposed a method to accurately classify polycystic ovaries using ultrasound images. The authors employed a combination of feature extraction techniques, specifically the Wavelet method, and a Convolutional Neural Network (CNN) to identify the unique characteristics of the ultrasound data. The results of the system testing indicated that the CNN achieved the highest accuracy of 80.84% in accurately identifying the polycystic ovaries.

AUTHORS	OBJECTIVES	RESEARCH DESIGN	RESULTS
S. Sreejith et al.[2022][4]	In order to help physicians monitor Polycystic Ovarian Syndrome, this study creates a clinical decision support system (PCOS).	Utilized a random forest classifier to analyze the characteristics after using the red deer method to identify the best ones.	In terms of accuracy, sensitivity, and specificity, the proposed methodology (RF+RDA) performs better than existing wrapper approaches employing RF and conventional classifiers, with scores of 89.81% accuracy, 90.43% specificity, and 89.73% sensitivity.
M A Anusuya et al.[2020][5]	A method for assessing and tracking symptoms that allows the chance of having PCOS to be predicted based on characteristics like testosterone levels, hirsutism, family history, Obesity, etc.	KNN, Linear Regression, and Random Forest are some of the machine learning supervised classification algorithms that have been utilized for prediction tasks.	The random forest approach outperforms the other two algorithms by averaging lower error levels (average MAE and RMSE values of 1.99 and 3.10, respectively) and the highest R^2 values (average 0.985).
B Rachana et al.[2021][6]	Discovering a way to detect PCOS in its earliest stages to avert additional difficulties.	The suggested technique includes a KNN classifier, which is primarily focused on decreasing a number of flaws, and classification is done using the KNN algorithm.	It is feasible to demonstrate that the KNN classifier has an accuracy of nearly 97%, which is higher than any classifier that has previously been suggested.
Vaidehi Thakre et .al.[2020] [7]	A method that can anticipate the therapy for PCOS based on an ideal and minimum set of characteristics has been presented.	Random forest, SVM, Logistic Regression, Gaussian Naive Bayes, and KNN are the five algorithms that were tested to predict PCOS.	Of the four, the Random Forest Classifier was found to be the most trustworthy and accurate, with an accuracy rate of 90.9%.
A.K.M. Salman Hosain et al.[2022][8]	To detect Polycystic Ovary Syndrome(PCOS) using Convolutional Neural Network Architecture from Ovarian Ultrasound Images	For the purpose of classifying data, they had created the pre-trained model InceptionV3 and the CNN model PCONet to identify ovarian cysts in ultrasound images.	The best model created is PCONett, which has an accuracy rate of 93.93%.
Amsy Denny et al.[2019][9]	Machine Learning-Based Diagnosis and Prediction System For Polycystic Ovary Syndrome (PCOS)	Used Logistic regression and six other algorithms such as Linear Discriminate Analysis, KNN, CART, RFC, NBC, SVM.	The best performance was given by Random Forest Classifier model, where an accuracy of 89 % was achieved after data optimization.
Kinjal Raut et al.[2022][10]	Machine Learning	Decision Tree, SVC, Random	Comparing CatBoostClassifier to other models, it

	Algorithms for PCOS Detection.	Forest, Logistic Regression, K Nearest Neighbor, XGBRF, and CatBoost Classifier are the methods used to build the model.	has excelled and achieved the greatest accuracy of 94.64%.

Table-1: Summary of Literature Review

III. METHODOLOGY

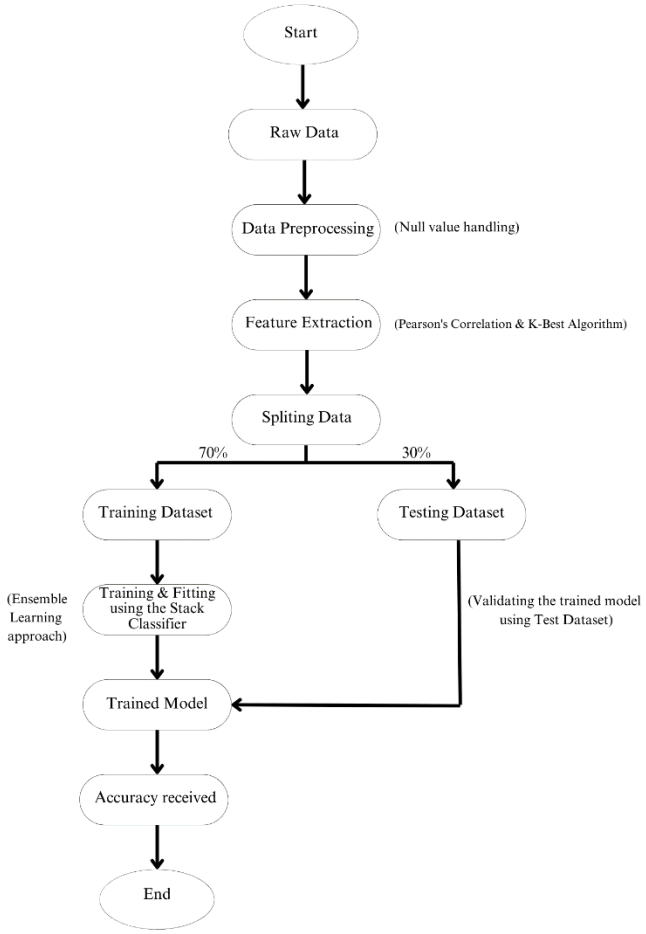


Fig-1: The flowchart of the entire process, from data gathering to precision getting

Proposed Algorithm: -

Step 1: *Dataset Description: -*

Polycystic ovary syndrome (PCOS) is a condition characterized by menstrual irregularities and high levels of male hormones. It is a significant contributor to infertility in women. To develop an accurate diagnostic model for PCOS, a combination of clinical and non-clinical data is needed. The dataset used in this study comprises data from 541 patients with and without fertility issues who were diagnosed with PCOS. The data was collected from 10 different hospitals in Kerala, India and is available in the Kaggle database. Table 1 provides a detailed description of the parameters that are included in the dataset.

Sl. No.	Parameter name	Description
1	Age	Patient's age in years
2	Weight	Patient's weight in kg
3	Height	Patient's height in cm
4	BMI	Body mass index
5	Blood group	Blood group
6	Pulse Rate	Pulse rate in beats per minute
7	RR	Respiratory rate in breaths per minute
8	Hb	Haemoglobin counts in grams per decilitre
9	Cycle (R/I)	Whether cycle is regular (2) or not(4)
10	Cycle length(days)	Number of days of menstruation
11	Marriage Status (Yrs.)	Number of years since marriage
12	Pregnant(Y/N)	Whether pregnant (1) or not(0)
13	No. of abortions	Number of abortions
14	I beta-HCG(mIU/mL)	Amount of beta human chorionic gonadotropin
15	II beta-HCG(mIU/mL)	Amount of beta human chorionic gonadotropin
16	FSH(mIU/mL)	Amount of follicles stimulating hormone
17	LH(mIU/mL)	Amount of Luteinizing hormone
18	FSH/LH	Ratio of FSH to LH
19	Hip(inch)	Hip size in inches
20	Waist(inch)	Waist size in inches
21	Waist: Hip Ratio	Waist to hip ratio
22	TSH (mIU/L)	Amount of Thyroid Stimulating hormone
23	AMH (ng/mL)	Amount of Anti Mullerian hormone
24	PRL (ng/mL)	Amount of Prolactin
25	Vit D3 (ng/mL)	Amount of Vitamin D3
26	PRG (ng/mL)	Amount of progesterone
27	RBS (mg/dl)	Random Blood Glucose
28	Weight gain(Y/N)	Whether the patient gained weight (1) or not (0)
29	hair growth(Y/N)	Whether the patient had hair growth (1) or not (0)
30	Skin darkening (Y/N)	Whether the patient had skin darkening (1) or not (0)
31	Hair loss(Y/N)	Whether the patient experienced hair loss (1) or not (0)
32	Pimples(Y/N)	Whether the patient has pimples (1) or not (0)
33	Fast food (Y/N)	Whether the patient consumes fast food (1) or not (0)

34	Reg..Exercise(Y/N)	Whether the patient exercises regularly (1) or not(0)
35	BP_Systolic (mmHg)	Systolic pressure
36	BP_Diastolic (mmHg)	Diastolic pressure
37	Follicle No. (L)	No: of follicles in the left ovary
38	Follicle No. (R)	No: of follicles in the right ovary
39	Avg. F size (L) (mm)	Average size of follicles in the left ovary
40	Avg. F size (R) (mm)	Average size of follicles in the left ovary
41	Endometrium (mm)	Thickness of the endometrium
42	PCOS(Y/N)	Diagnosed with PCOS (1) or not(0)

Table-2: A complete list of the dataset's characteristics

The dataset comprises numerical and categorical data, with physical parameters including age, weight, height, BMI, waist and hip dimensions, hair growth, hair loss, skin darkening, and pimples. The dataset also includes clinical parameters such as blood group, Vitamin D3 levels, pulse rate, respiration rate, hemoglobin count, cycle regularity, glucose levels, hormone levels, blood pressure, follicle count, follicle size, and endometrial thickness.

Step 2: *Data Preprocessing: -*

Preparing data for machine learning requires dealing with missing data, categorical variables, scaling features, and selecting important features. In this study, missing values in the dataset were replaced with 0 to ensure that the model can process the data. Before being fed into the model, samples with missing values were either removed or replaced with pre-built estimators.

Step 3: *Feature Selection: -*

Feature selection is a crucial step in building a ML model, as it can improve its performance by removing irrelevant features and reducing data dimensionality and algorithmic difficulty. The K-best algorithm selects top k features from a dataset based on their statistical scores. It's a filter-based approach that uses statistical tests to rank each feature. Top k features with highest scores are selected and rest are removed, reducing data dimensionality, and improving machine learning model efficiency and accuracy. Various methods such as Pearson's correlation coefficient, Chi-square test, mutual information, and Fisher's test can be used to evaluate the relationship between each input feature and the class variable to select the features that exhibit a strong relation. The best 20 features in this research were chosen using the K-best algorithm and Pearson's correlation coefficient approach.

1. Pearson's Correlation approach: -

Pearson correlation is a measure of the linear correlation between two variables. It is commonly used in machine learning to determine the strength and direction of the relationship between two numerical variables. The Pearson correlation coefficient, denoted as "r", ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

Pearson correlation is often used in feature selection, where the correlation between each feature and the target variable is calculated and features with a low correlation are removed from the dataset. In this study the correlated features like BMI, FSH/LH, Waist(inch) are dropped after identifying using Pearson's Correlation approach.

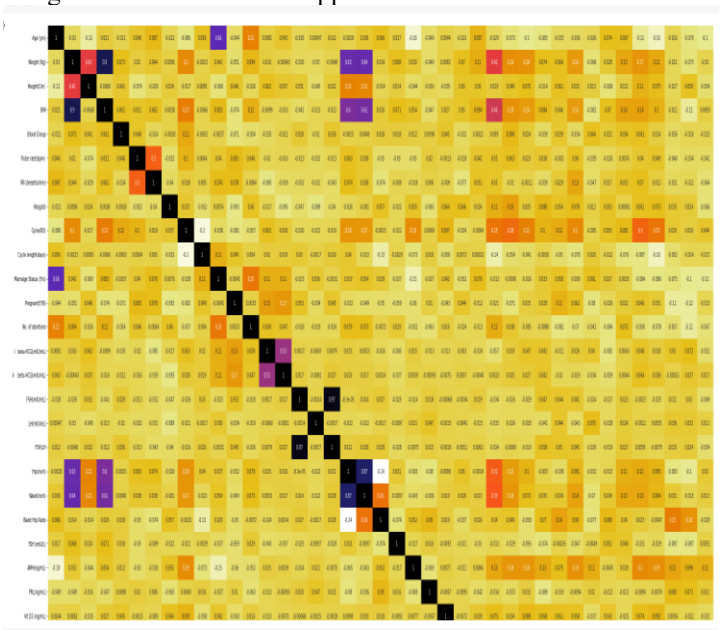


Fig-2: The dataset's entire feature correlation matrix is displayed in this image. Features that correlate most strongly are depicted by dark colours, while those that correlate least strongly are depicted by pale colours.

2. K-Best Algorithm approach: -

The K-best algorithm is a feature selection method in machine learning that selects the K best features from a larger set of features. This algorithm ranks the features based on their importance scores and selects the top K features with the highest scores.

The importance scores of the features are typically calculated using statistical methods such as mutual information, correlation coefficient, or chi-squared test.

	Features	**Score**
24	Vit D3 (ng/mL)	9477.648952
13	I beta-HCG(mIU/mL)	6950.525631
16	LH(mIU/mL)	2558.471157
15	FSH(mIU/mL)	1601.143311
14	II beta-HCG(mIU/mL)	949.362075
37	Follicle No. (R)	672.789402
36	Follicle No. (L)	573.647927
22	AMH(ng/mL)	233.210799
17	FSH/LH	96.831682
29	Skin darkening (Y/N)	84.870716
28	Hair growth(Y/N)	84.854623
27	Weight gain(Y/N)	65.554147
1	Weight (Kg)	49.466423
32	Fast food (Y/N)	37.721883

8	Cycle(R/I)	27.681419
25	PRG(ng/mL)	24.638020
31	Pimples(Y/N)	22.587803
10	Marraige Status (Yrs)	22.181398
3	BMI	14.568227
0	Age (yrs)	14.284370
30	Hair loss(Y/N)	8.846546

Table-3: Top 21 features in the collection, ranked and scored using the K-best feature selection algorithm

PCOS Dataset	Features
Total no. of Features	44
Selected no. of features and their names	21 Weight (Kg), Cycle(R/I),I beta-HCG(mIU/mL), II beta-HCG(mIU/mL),FSH(mIU/mL),LH(mIU/mL),FSH/LH,AMH(ng/mL),Vit D3 (ng/mL),PRG(ng/mL), Weight gain, hair growth(Y/N),Skin darkening (Y/N), Hair loss(Y/N),Pimples(Y/N),Fast food (Y/N),Follicle No. (L),Follicle No. (R),Avg. F size (L) (mm),Avg. F size (R) (mm),Endometrium (mm)

Table-4: Collection of finalized feature values for machine learning models

Step 4: *Training and testing dataset splitting:*

Splitting the pre-processed dataset into training and testing sets is a standard practice to evaluate the predictive model's performance. The training set is used to train and tune the model, while the test set is kept aside as "new" data to evaluate the model's prediction ability on unseen data. The model's performance is validated using cross-validation on the training set.

Training Data	Testing Data
70%	30%

Table-5: Training & Testing dataset splitting

Step 5: *Model Selection: -*

A study was conducted to establish a baseline using a selected set of features in several classifier algorithms. From the vast number of existing machine learning algorithms, only those that have been demonstrated to provide the best results in detecting PCOS and non-PCOS conditions are utilized and listed below.

1) Random Forest Classifier
2) Support Vector Machine
3) Stack Model (Ensemble approach of four ML Models)
4) Naïve Bayes Classifier
5) Logistic Regression (LR)
6) K-nearest neighbors (KNN)

Classifiers	Accuracy on Training Dataset	Accuracy on Testing Dataset
Random Forest	100%	91%
Support Vector Machine	92%	90%
Stack Model (RF + SVM + Naïve Bayes + Logistic Regression)	**97%**	**90%**
Logistic Regression	86%	88%
K-Nearest Neighbor	80%	69%
Naïve Bayes	85%	86%

Table-6: A comparison accuracy chart between our suggested stack model and the other ml algorithms

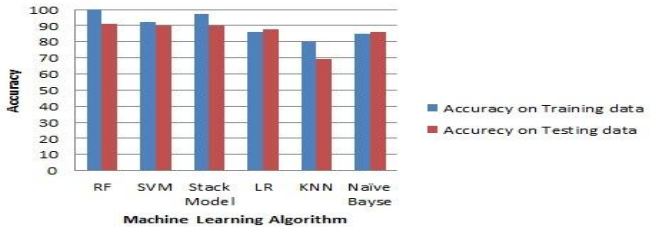


Fig-3: Accuracy graph of various classifiers along with proposed model

Proposed Stack Model (Four ML models are learned collectively for increased precision): -

In machine learning, a stack model, also known as a stacked generalization or stacked ensemble, is a technique that combines multiple predictive models to improve overall accuracy and reduce variance. The basic idea behind a stack model is to train several individual models on the same dataset, and then use a meta-model to combine the predictions of the individual models. The meta-model can be trained on the same dataset, using the predictions of the individual models as input features, or it can be trained on a separate holdout dataset. To improve accuracy, we have combined four algorithms (Random Forest, Naïve bayes, Support Vector Machine and Logistic Regression) in this research.

Highest Accuracy: -

With accuracies of 91%, 90%, and 90% on the test dataset, the top three algorithms are Random Forest, SVM, and our proposed Stack model (Random Forest + Support Vector Machine + Naïve Bayes + Logistic Regression).

When it comes to accuracy, Random Forest Classifier is the best.

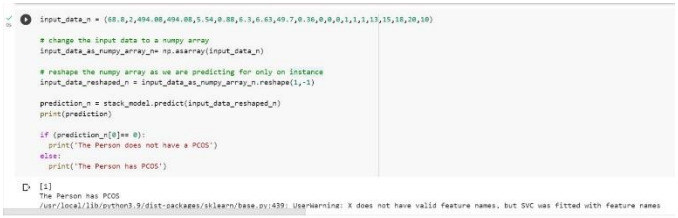
IV. RESULT & DISCUSSION

A total of 541 cases, which were gathered from different Thrissur infertility treatment facilities, were available for the research.

Accuracy score, confusion matrix, F1 score, precision, and recall are used to evaluate the performance of different models.

Algorithm used	Precision	Recall	F1-score	Support
Random Forest	0.93	0.91	0.92	163
Logistic Regression	0.90	0.89	0.89	163
Support Vector Machine (SVM)	0.91	0.91	0.91	163
K Nearest Neighbor	0.77	0.69	0.72	163
Naïve Byers	0.90	0.87	0.87	163

Table-7: Precision,F1 score and recall of different models along with proposed model



Output Screenshot-1: Prediction result of Stack model

CONCLUSION

To create awareness among the women we decided to create a prediction model to detect PCOS in its early stages. We developed a model using a Kaggle dataset with 20 features to detect PCOS in its early stages with 91% accuracy. The system we have developed can help doctors identify potential patients and give PCOS patients priority. In the future, it will be possible to use CNN to identify ovarian cancer in women with PCOS, who have a higher risk of developing the disease. The results of this study have substantial ramifications for improving early detection and treatment of PCOS, which can have detrimental effects on women's health and wellbeing.

REFERENCES

[1] Sumathi, M., Chitra, P., Sakthi Prabha, R., & Srilatha, K., "Study and detection of PCOS related diseases using CNN", IOP Conference Series: Materials Science and Engineering, vol. 1070, 2021.

[2] Talib, I., Khadija, S., Khan, A. M., Akram, S., Akhtar, M. K., Willayat, F., & Iftikhar, A., "Prediction a woman having Polycystic Ovary Syndrome (PCOS) those having Insulin Resistance (IR)", Pakistan Journal of Medical and Health Sciences, vol. 16, no. 2, pp. 6–9, 2022.

[3] Dewi, R & Adiwijaya, Kang & Wisesty, Untari Novia & Jondri, "Classification of polycystic ovary based on ultrasound images using competitive neural network", Journal of Physics: Conference Series, vol. 971, 2018.

[4] Sreejith, S., Khanna Nehemiah, H., & Kannan, A., "A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier", Healthcare Analytics, vol. 2, 2022.

[5] Pushkarini, H., & Anusuya, M. A., "A prediction model for evaluating the risk of developing PCOS", Journal of Medical Systems, vol. 44, no. 3, pp. 1-9, 2020.

[6] B Rachana., Priyanka, T., Sahana, K. N., Supritha, T. R., Parameshachari, B. D., & Sunitha, R., "Detection of polycystic ovarian syndrome using follicle recognition technique", Global Transitions Proceedings, vol. 2, no. 2, pp. 304-308, 2021.

[7] Vedpathak, S. & Thakre, V., "PCOcare: PCOS Detection and Prediction using Machine Learning Algorithms", Bioscience Biotechnology Research Communications, vol. 13, pp. 240-244, 2020

[8] Salman Hosain, A.K.M., Mehedi, M.H. and Kabir, I.E., "PCONet:A convolutional neural network architecture to detect polycystic ovary syndrome (PCOS) from ovarian ultrasound images", International Conference on Engineering and Emerging Technologies (ICEET), 2022.

[9] Denny, A. & Raj, A. & Ashok, A. & Ram, M.& George, R., "i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques", pp. 673-678,2019.

[10] Raut, K., Katkar, C., & Itkar, S. A., "PCOS Detect using Machine Learning Algorithms", International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 6, pp. 1376-1381,2020.