Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2-January 2024
Volume7-Issue2-January 2024
Paper: 39

Analysis of Polycystic Ovary Syndrome Using Machine Learning Algorithms

Anu Srishti Bara[1], Bhavya Jha[2], Dr. B. Arunsundar[3],

[1,2] Students,and [3] Faculty
Dept. of Data Science and Business Systems,
SRM Institute of Science and Technology,
Chennai, India.
ab5120@srmist.edu.in, bj7528@srmist.edu.in

Abstract: *This paper focuses on the data-driven opinion of polycystic ovary pattern (PCOS) in women. For this, machine literacy algorithms are applied to a dataset freely available in the Kaggle depository. This dataset has 43 attributes of 541 women, among which 177 are cases of PCOS complaints. Originally, a univariate point selection algorithm was applied to find the stylish features that can prognosticate PCOS. The ranking of the at- tributes is reckoned, and it's set up that the most important trait is the rate of follicle-stimulating hormone (FSH) and luteinizing hormone (LH). Next, holdout and cross-confirmation styles are applied to the dataset to separate the training and testing data. A number of classifiers similar to grade boosting, arbitrary timber, logistic retrogression, and cold-blooded arbitrary timber and logistic retrogression (RFLR) are applied to the dataset. Results show that the first 10 loftiest-ranked attributions are good enough to predict the PCOS complaint. Results also demonstrate that RFLR exhibits stylish testing delicacy and recall value-important features. Hence, RFLR is suitable for reliably distinguishing PCOS cases.*

Keywords: *Random Forest, Decision Tree, Ada Boost, XGBoost and Hybrid model*

1. INTRODUCTION:

Polycystic ovary syndrome (PCOS) is a regular endocrine syndrome in females affected by elevated androgen. Treat- ment of PCOS is based on some cardinal features including anovulation, hyperandrogenism signs, and menstrual dysfunc- tion. Some of the important symptoms of this disorder are pelvic pain, excess hair, acne, hirsutism, velvety skin, male hormone, irregular periods, danker, etc. Approximately 5- 10% of reproductive age (15-49 years) women suffer from this problem. The outbreak of PCOS was documented to be 8% and 4.8% in African American women and white women, respectively [1-2]. Women with this ovarian dys- function are associated with hypertension, increased risk of cardiovascular disease, obesity, gynecological cancer, type 2 diabetes mellitus, etc. Moreover, recent research has shown that PCOS causes higher risk of first trimester miscarriage. PCOS causes in ovaries inappropriate growth of follicles that are prevented at a primary stage and miscarry to mature. This is one of the causes for infertility. Therefore, it is significant to screen the patients at a primary stage to prevent any serious moment of the PCOS disease. [1] This paper focuses on the analysis of PCOS. The main contributions of this work are as follows: (i) selecting the most important attributes of PCOS patients for the given dataset using feature selection method. (ii) Applying machine learning algorithms on the important features of the PCOS dataset. (iii) Comparing the machine learning algorithms with those reported in the literature in terms of testing accuracy and recall.

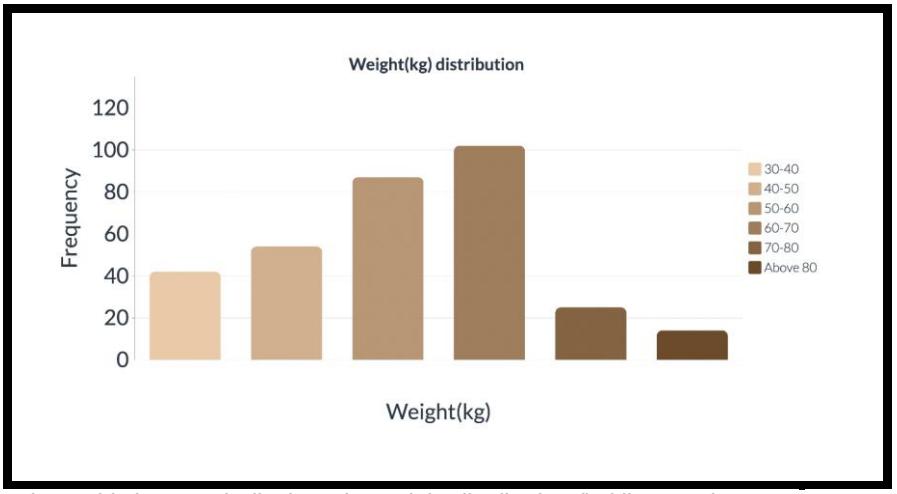


Fig. 1.This bar graph displays the weight distribution (in kilograms) among women suffering from PCOS (Polycystic Ovary Syndrome). The majority of women fall within the weight range of 50-70 kg, with 60- 70 kg being the most prevalent category.

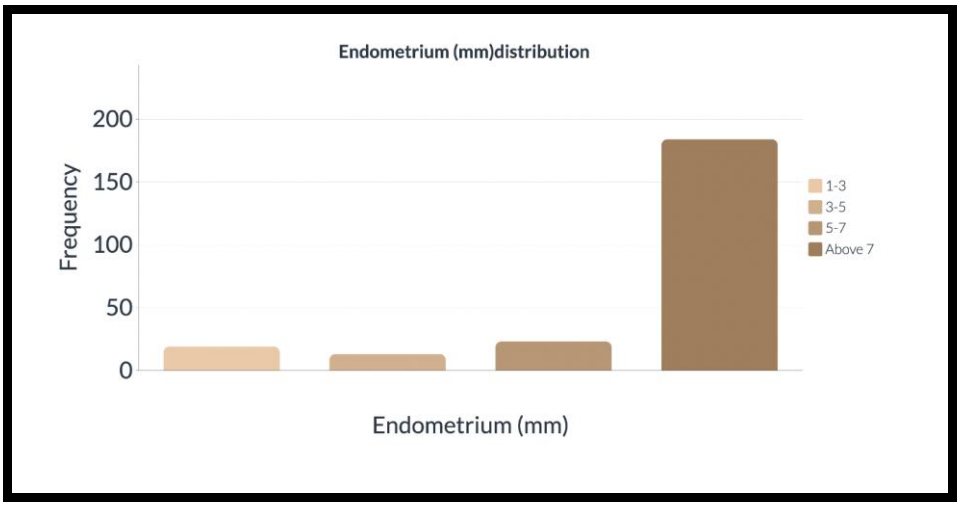


Fig. 2.The bar graph displays the pattern of the thickness of the endometrium (in millimeters) in individuals with Polycystic ovary syndrome (PCOS). Women diagnosed with polycystic ovarian syndrome have been seen to have thick endometrial linings.

2. LITERATURE REVIEW

A. Samia Ahmed et al., "A Review on detection techniques of PCOS using Machine Learning techniques" [2023]

Provided a comprehensive overview of existing research methodologies and algorithms utilized in the detection and classification of polycystic ovary syndrome (PCOS), includ- ing their performance metrics and limitations. The literature review reveals a diverse array of approaches and algorithms employed in PCOS detection. Despite variations in method- ologies, studies consistently strive for improved accuracy rates, with some achieving remarkable results using advanced methods such as convolutional neural networks (CNNs) and ensemble learning techniques.

B. Rachana. B et al., "Detection of Polycystic Ovarian Syn- drome Using Follicle Recognition Technique" [2021]

The approach taken to detect PCOS is using ultrasound images and studying various features by combining segmentation, feature extraction, and the classification process. The various approaches for segmentation and classification were studied to enhance past research and obtain a model with greater accuracy. For classification, the KNN method is used.

C. J. Madhumitha et al., "Automated Polycystic Ovarian Syndrome Identification with Follicle Recognition" [2021]

In this paper, an automated system for detecting polycystic ovarian syndrome using follicle recognition and classification using three types of machine learning algorithm methods is compared. The proposed method achieves an accuracy of about 98%, suggesting its effectiveness.

D. Sayma Alma Suha et al., "extended machine learning techniques for polycystic ovary syndrome detection using ovary ultrasound images" [2022]

To develop a method for efficiently detecting Polycystic Ovary Syndrome (PCOS) from ovarian ultrasound images by integrating traditional machine learning and deep learning techniques. The study proposes a hybrid approach combining transfer learning with convolutional neural networks (CNNs) for feature extraction and stacked ensemble machine learning for classification.

E. E. Setiawati, Adiwijaya, T. A. B Wirayuda, and W. Astuti, "Categorization of PCOS Based on Ultrasound Images Us- ing Supervised Learning and Particle Swarm Optimization" [2011]

Further research on PCOS should investigate genetic and environmental factors influencing predisposition. The relation- ship between androgen excess and insulin resistance warrants re-evaluation, considering developmental perspectives. Quantifying contributions to steroidogenesis from ovarian, adrenal, and extra glandular sources is essential, along with estab- lishing universal reference levels. Understanding intraovarian regulation of follicle development and mechanisms of arrest requires further elucidation. The long-term

consequences of PCOS, including type 2 diabetes and cardiovascular diseases, need thorough investigation. Identifying susceptible individu- als using genomic and proteomic approaches can personalize therapy and prevention strategies. It's acknowledged that the review selectively focused on controversial areas, suggest.

F. Adiwijaya, B. Purnama, A. Hasyim, M. D. Septiani, U.N. Wisesty and W. Astuti, "Follicle detection on the images to support determination of polycystic ovary syndrome" [2015]

PCOS, an endocrine abnormality in the female reproductive cycle, is often detected through stereology, feature extrac- tion, and classification methods. This study utilizes a Gabor wavelet for feature extraction and a modified backpropagation algorithm for classification. The modified backpropagation incorporates Levenberg-Marquardt optimization and conjugate gradient-Fletcher reeves to enhance convergence rates. Levenberg-Marquardt optimization yields higher accu- racy (93.925%) but with longer running times, while conjugate gradientFletcher Reeves achieves 87.85% accuracy. Notably, the Levenberg-Marquardt optimization with 33 neutrons and 16 vector features produces the best accuracy.

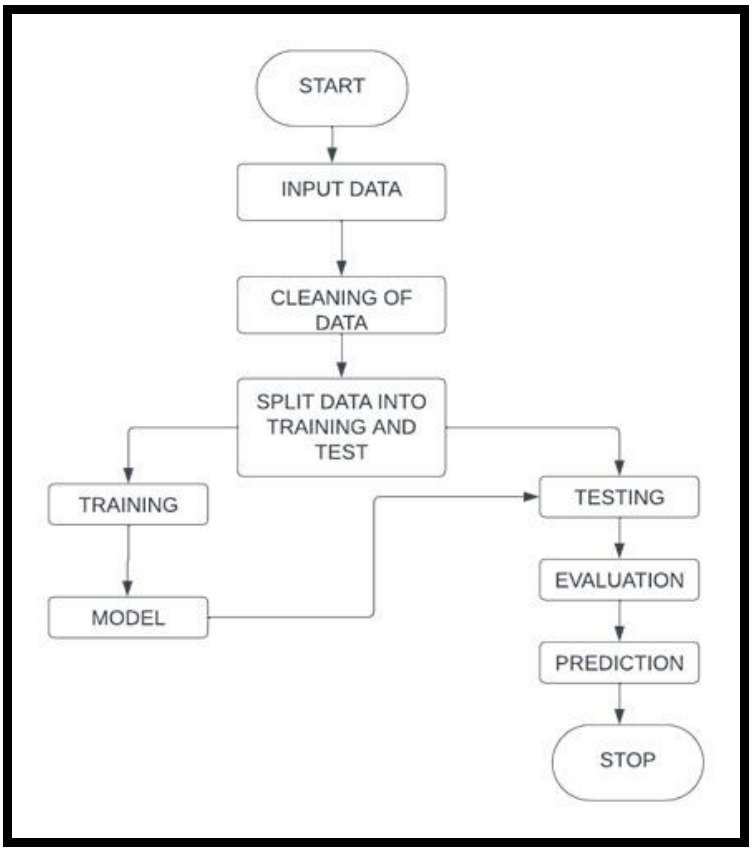
G. E. Setiawati, Adiwijaya, and A. Tjokorda, "Particle swarm optimization on follicle decomposition to support PCOS detec- tion." [2015]

PCOS, a prevalent endocrine disorder in female reproductive cycles, often manifests as polycystic ovaries (PCO), characterized by numerous small cysts or follicles. This study introduces an image clustering method for follicle segmenta- tion using Particle Swarm Optimization (PSO) with a novel modified fitness function incorporating the Mean Structural Similarity Index (MSSIM) and Normalized Mean Square Error (NMSE). Compared to previous approaches, the proposed PSO fitness function demonstrates enhanced convergence, particularly on ultrasound images. Additionally, the study explores the impact of contrast enhancement on PSO image clustering performance and follicular size extraction, showing that contrast enhanced PSO clustering yields closer Regions of Interest (ROI) to manually identified reference ROIs by medical professionals.

3. METHODOLOGY

Several machine learning models have been proposed to classify fraud as fraud or not fraud, yet they often fail to ad- dress misdiagnosis adequately. Similar challenges exist in diagnosing Polycystic Ovary Syndrome (PCOS) and evaluating tumors, where models overlook data heterogeneity and size. To tackle these issues, we propose a machine learning-based approach that incorporates novel data preprocessing techniques for feature transformation. By leveraging Random Forest, Decision Tree, AdaBoost, XGBoost, and a hybrid model, we aim to enhance accuracy while mitigating bias, instability, and deviation. Our methodology includes comprehensive classifier tests to validate the effectiveness of these techniques. Some of the benefits of the proposed methodology are:

- Requires less time
- Good accuracy and efficiency
- Easy to handle



A. ALGORITHMS

1) RANDOM TREE: A random forest is a machine learning approach that's used to break down retrogression and category challenges. It utilizes ensemble knowledge, which is a strategy that combines multiple classifiers to give answers to complicated problems. A random forest algorithm consists of multiple decision trees. The 'forest' generated by the random forest algorithm is seasoned through bagging or bootstrap aggregat- ing. Bagging is an ensemble meta-algorithm that improves the preciseness of machine learning algorithms. The random forest algorithm establishes a conclusion predicated on the prognosti- cations of the decision

trees. It predicts by taking the normal or mean of the yield from varied trees. Expanding the number of trees increases the accuracy of the resultant. A random forest eradicates the extent of a decision tree algorithm. It reduces the overfitting of datasets and increases perfection. It generates prognostications without taking numerous configurations in packages (like Scikit Learn). Features of a Random Forest Algorithm:

- It's more precise than the decision tree algorithm.
- It provides an efficient way of managing missing data.
- It can produce a reasonable prediction without hyperactive parameter tuning.
- It solves the effect of over serving in decision trees.
- In every random forest tree, a subset of features is opted aimlessly at the knot's splitting point.[10]

2) DECISION TREE: Decision tree learning is a supervised machine learning method for changing a decision tree from training data. A decision tree (also pertain to as a classi- fication tree or a reduction tree) is a predictive model which is a mapping from observances about an item to conclusions about its target value. The aim of using a Decision Tree is to generate a training model that can use to forecast the class or value of the target variable by learning simple decision rules inferred from previous data (training data). In Decision Trees, for prognosticating a class label for a record we start from the root of the tree. Decision trees are the building sets of a random forest algorithm. A decision tree is a decision support strategy that forms a tree- alike structure. An overview of decision trees will help us understand how random forest algorithms result. A decision tree consists of three elements decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further insulate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be separated further. The bumps in the decision tree depict attributes that are used for forecasting the result. Decision nodes supply a link to the leaves. [12] A decision tree is a versatile tool for classification and regression tasks, facilitating intuitive decision-making through a hierarchical structure of conditions. Its branching logic enables effective analysis of data, aiding in understanding complex relationships and guiding informed decisions in various domains.

3) XG BOOST: XGBoost stands for "Extreme Gradient Boosting". XGBoost is an optimized distributed grade boosting library aimed to be largely productive, adaptable and movable. It implements Machine Learning algorithms under the grade Boosting frame. It provides a resemblant tree boost- ing to break numerous data wisdom problems in a fast and accurate way. Boosting is an ensemble learning approach to make a strong classifier from several weak classifiers in series. Boosting algorithms play a pivotal part in dealing with bias- friction trade-off. Unlike bagging algorithms, which only control for high friction in a model, boosting controls both the aspects (bias and friction) and is considered to be more effective.

4) ADA BOOST: AdaBoost algorithm, short for Adaptive Boosting, is a Boosting approach used as an Ensemble Method in Machine Learning. It's called Adaptive Boosting as the weights are-assigned to each case, with advanced weights assigned to inaptly classified cases. Boosting is used to reduce bias as well as friction for supervised literacy. It works on the principle of learners growing successionally. Except for the first, each posterior learner is grown from preliminarily grown learners. In simple terms, weak learners are converted into strong learners. The AdaBoost algorithm works on the same principle as boosting with a little dissimilarity. AdaBoost works by iteratively training a sequence of weak learners, where each learner focuses on the cases that the former learners stumbled with. It assigns advanced weights to mis-

classified cases, so posterior weak learners prioritize getting those cases right. These weak learners are generally simple decision trees, frequently called" wholes," which only make opinions grounded on a single point. After each replication, AdaBoost adjusts the weights of the training cases grounded on the delicacy of the weak learner. Cases that were misclassified get advanced weights, and rightly classified cases get lower weights. This adaptation emphasizes the significance of the misclassified cases in posterior duplications. In the end, AdaBoost combines all the weak learners into a single strong learner by giving each weak learner a weight commensurable to its delicacy. This way, the final model is a weighted sum of the weak learners' prognostications, where more accurate weak learners have a lesser influence on the final decision. AdaBoost is a important algorithm for both bracket and retrogression tasks and is known for its capability to achieve high delicacy with fairly simple weak learners. still, it can be sensitive to noisy data and outliers.

5) HYBRID MODEL: A hybrid model in machine learn- ing combines multiple algorithms or techniques to improve overall performance. By leveraging the strengths of different approaches, it aims to achieve better predictive accuracy and robustness. This integration can involve ensemble methods, feature engineering, or domain-specific knowledge fusion.

4. RESULT AND DISCUSSIONS

A. SYSTEM REQUIREMENTS:

1) Functional and non-functional conditions: Functional and non-functional conditions demand's analysis is a veritably critical process that enables the success of a system or software design to be assessed. Conditions are generally resolved into two types, Functional and nonfunctional conditions.

2) Functional conditions: : These are the requirements that the end-user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of input to be given to the system, the operation performed and the output expected. They are basically the requirements stated by the end user which one can see directly in the final product, unlike the non-functional requirements.

Examples of functional conditions:

- Authentication of the end user whenever he she logs into the system.
- System arrestment in case of cyber-attack.
- A verification dispatch is transferred to the end-user when- ever he/ she registers for the first time on some software system.

3) Non-functional conditions: : These are principally the quality constraints that the system must satisfy according to the design contract. The precedence or extent to which these factors are enforced varies from one design to another. They're also called non-behavioral conditions.

Applications of non-functional conditions:

- Emails should be transferred with a quiescence of no lesser than 12 hours from such an exertion.
- The processing of each request should be done within 10 seconds.

3) H/W Configuration:

H/W Configuration	
Processor	I3/ Intel Processor
Hard Disk	160 GB
RAM	8 GB

4) S/W Configuration:

S/W Configuration	
Operating Systems	Windows 7/8/10
Server-Side Script	HTML, CSS, JS
IDE	PyCharm
Libraries Used	NumPy, IO, OS, Flask, Keras
Technology	Python 3.6+

Properties	*Comparison Table*				
	Random Forest	*Decision Tree*	*XG Boost*	*AdaBoost*	*Hybrid*
Accuracy	*82%*	*82%*	*88%*	*86%*	*89%*
Time Complexity	*O(n*m* log(m))*	*O(n*m*k* log(m))*	*O(n*m*T)*	*O(n*m*T)*	*Depends on Specific Combination*
Objective	*To improve prediction accuracy and reduce overfitting.*	To iteratively optimize an ensemble of weak learners to minimize a predefined loss function.	*To sequentially train weak learners and combine them into a strong learner.*	*To create a model that predicts the value of a target variable by learning simple decision.*	*To leverage the strengths of different algorithms or techniques to improve overall prediction.*

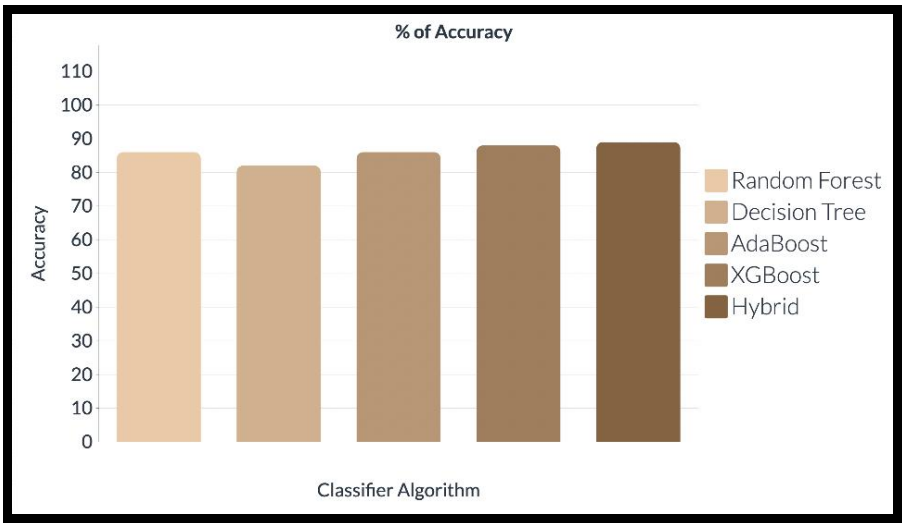


Fig. 3.This bar graph illustrates the percentage accuracy of various classified algorithms. Hybrid models lead with the highest accuracy, followed closely by Random Forest and XGboost. Overall, Hybrid models emerged as the top performer among the showcased algorithms

5. CONCLUSION

In conclusion, our project introduces a user-friendly application designed for diagnosing Polycystic Ovary Syndrome (PCOS) using state-of-the-art machine learning algorithms. By employing Random Forest, Decision Tree, Ada Boost, XGBoost, and a Hybrid model, we have leveraged the most effective techniques available in the field. Through rigorous development and testing, our application accurately determines whether an individual is likely to have PCOS or not. The culmination of our efforts results in a reliable tool that provides valuable insights into PCOS diagnosis. With its intuitive inter- face and robust algorithmic framework, our application offers a practical solution for healthcare professionals and individuals seeking timely and accurate diagnoses. Furthermore, our com- mitment to transparency and interpretability ensures that users can trust and understand the diagnostic outcomes produced by our application. Moving forward, ongoing refinement and validation will enhance the performance and reliability of our diagnostic tool, ultimately improving healthcare outcomes for those affected by PCOS.

REFERENCES

[1] R. Pasquali et al., “PCOS Forum: Research in Polycystic Ovary Syn- drome Today and Tomorrow”, Clin Endocrinol (Oxf), vol.74, no. 4, pp. 424–433, 2011. Doi: 10.1111/j.1365-2265.2010. 03956.x

[2] A. S. Lagana‘, S.G. Vitale, M. Noventa, and A. Vitagliano, “Cur- rent 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 determi- nationo f polycystic ovary syndrome”, 3rd International Conference on Science and Engineering in Mathematics, Chemistry and Physics 2015 (ScieTech2015), vol. 622, 2015.

[4] U. N. Wisesty, J. Nasri and Adiwijaya, “Modified backpropagation algorithm for PCOS detection based on ultrasound images”, Recent Advances on Soft Computing and Data Mining - The Second Interna- tional Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, pp.144-151, August 18-20, 2016.

[5] Adiwijaya,M.Maharani,B.Dewi,F.YuliantoandB.Purnama,“Digital- image compression using graph coloring quantization based on wavelet- svd”, 2013 International Conference on Science and Engineering in- Mathematics, Chemistry and Physics (ScieTech 2013), vol. 423, 2013.

[6] 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 images”, 3rd International Conference on (IEEE) Information and Communication Technology (ICoICT), pp 396-401,2015.

[7] E. Setiawati, Adiwijaya, T. A. B Wirayuda, W. Astuti, “A Classifica- tionof Polycystic Ovary Syndrome Based on Ultrasound Images Using Supervised Learning and Particle Swarm Optimization”, Advanced Science Letters, vol. 22, pp.1997-2001, 2016.

[8] P. Mehrotra, C. Chakraborty, B. Ghoshdastidar, S. Ghoshdastidar and K. Ghoshdastidar, “Automated ovarian follicle recognition for polycys- tic ovary syndrome”, International Conference on Image Information Processing (ICIIP), pp 1-4, 2011.

[9] S. Rihana, H. Moussallem et. al, “Automated algorithm for ovarian cysts diagnosis in ultrasonography”, 2nd International Conference on Advances in Biomedical Engineering (ICABME), pp 219-222, 2013.

[10] S. Ahmed et. al.,” A Review on Detection Techniques of Polycystic Ovary Pattern Using Machine literacy,” in IEEE Access, vol. 11, pp. 86522- 86543, 2023, doi10.1109/ACCESS.2023.3304536.

[11] P. Chauhan, P. Patil, N. Rane, P. Raundale and H. Kanakia, ”Com- parative Analysis of Machine Learning Algorithms for Prediction of PCOS,” 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 2021.

[12] N. Nabi, S. Islam, S. A. Khushbu and A. K. M. Masum, ”Machine Learning Approach: Detecting Polycystic Ovary Syndrome and It’s Impact on Bangladeshi Women,” 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021.

[13] A. Tanwar, A. Jain and A. Chauhan, ”Accessible Polycystic Ovarian Syndrome Diagnosis Using Machine Learning,” 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022.

[14] S. Aggarwal and K. Pandey, “Determining the representative features of PCOS via the Design of Experiments,” Multimed.Tools Appl., vol. 81, pp. 29207–29227, 2022.

[15] S. Łukasz and W. Jakub,” Toward automatic assessment of a threat of women’s health diseases grounded on ontology decision models and menstrual cycle analysis,” 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021