iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51426

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Detection of PCOS using Ensemble Models

Prof Ajil A[1], Tanvi Jain [2], T M Namratha [3], Vismaya S [4], Thummaluru Ganga Lakshmi[5]
Dept. of Computer Science and Engineering, REVA University, Bengaluru, India

Abstract: ***Polycystic ovary syndrome (PCOS) is a complex endocrine disorder that affects women of childbearing age. It is characterized by a range of symptoms, including irregular monthly cycles, hirsutism, and childlessness. Early diagnosis and detection of PCOS is vital for successful management of this condition. In the last few years, machine learning algorithms have shown great results in the diagnosis of various medical conditions. The proposed model is an ensemble model consisting of XG Boost and Random Forest to detect PCOS in women by analysing a dataset of 541 women, including 177 patients with PCOS. We analyse a dataset of clinical and demographic variables from women with and without PCOS and use various machine learning algorithms such as Ada boost, Random forest ,XG boost, Decision tree and a hybrid model to predict the presence of the condition. We evaluate the accuracy of our models by comparing the performance of the above listed algorithms. The proposed method consists of a hybrid model which is a combination of two algorithms that is Random forest and XGBoost and is yielding one of the highest accuracies of 97.2%.This early detection could potentially improve the level of care for women with this condition.***

Keywords: PCOS, Machine learning algorithms, Decision tree, Random forest, Ada boost, XG Boost.

I. INTRODUCTION

Polycystic ovarian syndrome (PCOS) is a prevalent endocrine disorder that affects women of childbearing age, with upto 20% of the population believed to be affected by this condition [2] It is characterized by a range of symptoms, including menstrual irregularities, acne, hirsutism, and infertility, among others.[3] Despite its high prevalence and significant impact on women's health, the diagnosis of PCOS is often challenging and relies on a combination of clinical and biochemical criteria.

Traditional diagnostic methods for PCOS, such as the Rotterdam criteria, are often cumbersome and expensive, and there is a need for a non-invasive diagnostic approach that can potentially facilitate earlier diagnosis and more effective treatment . In the last few years, machine learning algorithms have emerged as a one of the potentially effective tools for the early identification and diagnosis of PCOS. Machine learning algorithms are a sub-branch of artificial intelligence that empowers computers to discover and grasp the data and employ it to generate predictions or make informed choices. These algorithms have shown significant potential in a range of applications, from image recognition and natural language processing to healthcare and medical diagnosis.

In the context of PCOS, machine learning algorithms can be used to examine or look over large datasets of clinical and biochemical features to extract meaningful patterns and associations that may be useful for diagnosis and treatment. By training these algorithms on a dataset of patients with PCOS and a control group of patients without PCOS, we can develop models that can accurately and efficiently diagnose PCOS based on a range of features.

Utilizing machine learning algorithms for the diagnosis of PCOS has multiple advantages. First, it may enable earlier diagnosis of the disorder, which can lead to more effective treatment and better outcomes for patients. Second, it may provide a non-invasive and less expensive alternative to traditional diagnostic methods, which can be particularly beneficial in resource-limited settings. Finally, it may enable the identification of new biomarkers or features that are presently not incorporated into the PCOS diagnostic criteria. However, the use of machine learning algorithms for the diagnosis of PCOS also presents several obstacles and constraints . One of the main obstacle is the need for extensive and heterogenous datasets to train and validate these algorithms. This can be particularly challenging in the case of PCOS, which is a complex and heterogeneous disorder that may present differently in different populations.

Another obstacle is the need to ensure that these algorithms are not biased and do not perpetuate existing health disparities. The effectiveness of machine learning algorithms depends on the quality of the training data, and if the data is prejudiced or inadequate, the resulting models may also exhibit prejudice or inadequacy.

Finally, there is a need to ensure that these algorithms are transparent and explainable, so that clinicians and patients can understand how the diagnosis was reached and can trust the results. This is particularly important in the case of PCOS, where patients may be skeptical of a diagnosis based on machine learning algorithms alone.

Despite these difficulties, leveraging machine learning algorithms for PCOS diagnosis presents substantial potential and could revolutionize women's healthcare. This paper aims to examine the capability of machine learning algorithms to detect and diagnose PCOS early, utilizing a vast collection of clinical and biochemical characteristics.We will evaluate the performance of several algorithms, including Random Forest Classifier,Decision tree,XG Boost,Ada boost and compare their accuracy and efficiency. We will also perform feature selection to identify the most informative features for the diagnosis of PCOS and analyze the importance of these features using feature importance analysis. The outcomes of our study may have noteworthy ramifications for creating non-invasive diagnostic methods for not only PCOS but also other illnesses.

II. LITERATURE REVIEW

This paper by B. Purnama, U. N. Wisesti, Adiwijaya, F. Nhita, A. Gayatri, and T. Mutiah[1] presented a grouping method for diagnosing polycystic ovary syndrome (PCOS) based on follicle detection of ultrasound images .Their study achieved an accuracy of 82.55% using SVM-RBF Kernel on C=40. Conversely, our research used a dataset with more attributes and patients of PCOS disease, and achieved a higher accuracy of 93.45% using the RFLR algorithm. Additionally, we employed univariate feature selection to identify the most significant attribute and experimented with various classifiers including gradient boosting, random forest, and logistic regression.

The paper by P. Mehrotra et al. [2] proposes an automated screening method for PCOS using machine learning techniques. The authors employ a Bayesian classifier and utilize a dataset of clinical and demographic variables to predict the presence of PCOS. They report an accuracy of 93.93%, which suggests that their method is effective in the early detection of PCOS. This paper adds on to the existing literature on employing machine learning for PCOS diagnosis and accentuates the promise of Bayesian Classifier in this sphere.

In contrast to [2], this paper by Lawrence et al. [3] utilized a linear discriminant classifier to assist in ultrasound images depicting polycystic ovarian morphology. Their study achieved an accuracy of 92.86%, slightly lower than [2]. However, their method offers another technique for the automatic identification of PCOS using ultrasound pictures.

This paper by Cheng and Mahalingaiah[4] (2018) reported an accuracy of 97.6% using a rules-based classifier (RBC) for the categorization of polycystic ovaries using pelvic ultrasound examinations. Their study utilized data mining techniques to analyze the characteristics of ovaries, ultimately leading to the development of an RBC model.

This paper by , Dewi and Wisesty[5] proposed a classification method for polycystic ovary syndrome using competitive neural networks (C-NN). The suggested method obtained an accuracy rate of 80.84% in the categorization of ultrasound images.

This paper by Sachdeva et al.[6] focuses on the comparison of metabolic,clinical and hormonal variables between in women with PCOS who are non-obese and obese,as well as their distinct reaction to clomiphene. The research was carried out on a group of 80 females diagnosed with PCOS, including 40 who were obese and 40 who were not obese. The findings indicated that women with PCOS who were obese had elevated levels of luteinizing hormone (LH), testosterone, and insulin resistance in comparison to those who were not obese.

Additionally they noticed that obese females with PCOS demonstrated a reduced response to clomiphene when compared to non-obese females are non-obese. In conclusion,the research indicates that obesity can influence the metabolic,clinical and hormonal characteristics of PCOS and may impact the effectiveness of clomiphene treatment.

This paper by McCartney and Marshall's [7] provides a comprehensive overview of Polycystic Ovary Syndrome (PCOS), including its epidemiology, pathogenesis, clinical presentation, diagnosis, and management. They discuss the various diagnostic criteria and the role of imaging studies and laboratory testing in the diagnosis of PCOS. The article also emphasizes the correlation of PCOS with metabolic disruptions and an increased likelihood of cardiovascular disease, diabetes, and endometrial cancer. The authors underscore the significance of a comprehensive strategy in the treatment of PCOS, which includes lifestyle changes, pharmacological interventions, and reproductive assistance.

This paper published in Lancet[8]presents a thorough assessment of polycystic ovary syndrome (PCOS), emphasizing its description, diagnosis, and treatment. It examines the present knowledge of the underlying mechanisms of PCOS, its observable symptoms, and associated medical conditions. The researchers analyze different diagnostic standards utilized for PCOS and recommend the Rotterdam criteria as the most extensive approach.. The paper also reviews the different treatment options available for PCOS and highlights the need for a multidisciplinary approach to its management.

The paper by P.A. and Nestler [9] explores connection between metabolic syndrome and PCOS. The prevalence of the metabolic syndrome in PCOS is 2-fold higher than that for women in the general population.

The pathogenic link between the metabolic syndrome and PCOS is likely insulin resistance, and common metabolic abnormalities present in PCOS include obesity, atherogenic dyslipidemia, hypertension, impaired fasting glucose/impaired glucose tolerance, and vascular abnormalities. Lifestyle modification and pharmacological therapy with insulin-sensitizing agents are potential treatments for the metabolic syndrome in women with PCOS.

Women diagnosed with PCOS are twice as likely to have metabolic syndrome compared to the general population. Insulin resistance is considered the probable cause of the pathogenic connection between metabolic syndrome and PCOS, and common metabolic abnormalities linked with PCOS are obesity,hypertension, , impaired glucose tolerance/impaired fasting glucose, atherogenic dyslipidemia and vascular abnormalities. Changes to lifestyle and drug treatment with insulin-sensitizing agents are potential therapeutic options for metabolic syndrome in women with PCOS.

III. METHODOLOGY/ ALGORITHMS

The objective of this study is to explore the potential of machine learning techniques for the prompt detection of Polycystic Ovary Syndrome (PCOS), a prevalent hormonal disorder that impacts women in their childbearing years.There is a need for a non-invasive diagnostic technique because conventional PCOS diagnostic procedures can be time-consuming and expensive and may make it easier to diagnose the condition early and treat it more successfully.

With the aid of machine learning techniques, we will examine a sizable dataset of clinical and biochemical variables in this study in order to identify significant patterns and relationships. In particular, we will assess the performance of different algorithms, such as Ada boost,decision tree,Random Forest,XG Boost and hybrid algorithm to ascertain which method delivers the most precise and effective diagnosis of PCOS .

To do this, a dataset of patients with PCOS will be gathered, coupled with a control group of individuals without the condition. The data will then be cleaned, missing values will be imputed, and feature selection will be used to determine which features are most useful for the diagnosis of PCOS.

The selected machine learning techniques will next be trained and evaluated using the preprocessed dataset. Cross-validation will be used to make sure the models are reliable and aren't overfitting the training set of data. In the ultimate analysis, the performance of each algorithm will be compared to establish which approach provides the most accurate and efficient diagnosis of PCOS. In order to determine the most crucial characteristics for the PCOS diagnosis, we will also do feature importance analysis.

The primary objective of this research is to demonstrate the potential of machine learning techniques for early identification of PCOS, and to provide knowledge regarding the most effective methods and features for achieving this goal. The development of non-invasive diagnostic techniques for PCOS and other disorders may be significantly impacted by the findings of this study.

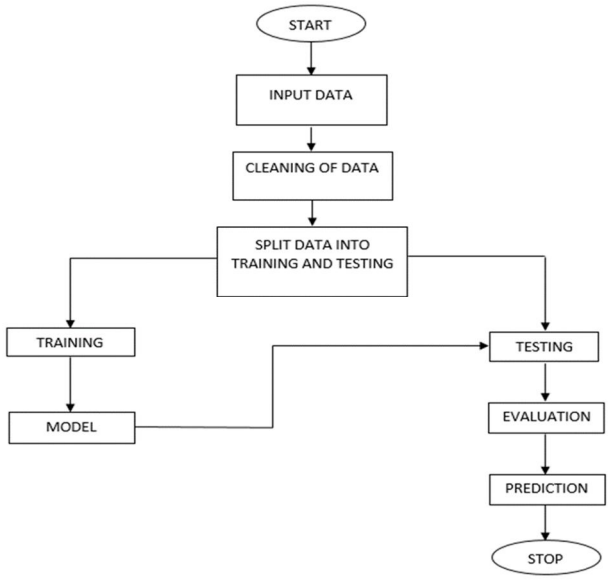


Fig 1 : diagram outlining of proposed method

A. *Algorithms Used*

1) *Decision Tree:* The decision tree is a frequently used supervised learning algorithm in machine learning that can be applied to classification and regression tasks. It constructs a model in the shape of a tree, where every node signifies an attribute/feature, and each branch indicates a decision or rule based on that feature. The decision tree is constructed by dividing the dataset into smaller subsets based on the most informative feature recursively until a specified stopping criterion is reached. The aim is to create a tree that can effectively predict the target variable or class for new data instances.Decision trees are easy to interpret and visualize, making them useful for understanding the underlying relationships in the data. They can also handle both categorical and numerical data, and are less affected by outliers compared to some other algorithms. However, they can suffer from overfitting if the tree is too complex or the data is noisy and may not be as effective as some other algorithms when working with extensive datasets or datasets containing highly correlated features. The Mathematical equation for Decision tree Algorithm is given(1)

$$E = -\sum_{i=1}^{n} p_i * \log(p_i) \quad (1)$$

where E is entropy , p_i is the probability of selecting an example from class i at random.

Random forest : It is an ensemble learning algorithm that employs numerous decision trees to enhance the accuracy and consistency of predictions. To build a set of decision trees, random forest algorithm randomly selects features and data points. Each tree in the forest is then trained on a random subset of the data. The algorithm aggregates the predictions of all trees in the forest to make the final prediction.This algorithm can handle both regression tasks and classification and is effective in handling high-dimensional and complex datasets. The use of multiple trees in random forest helps to reduce overfitting compared to a single decision tree, making it more reliable for making predictions on new data.However, random forest can be slower to train than some other algorithms due to the multiple trees it builds, and the resulting model may be harder to interpret compared to a single decision tree.The mathematical equation for Random forest algorithm that is the calculation of Gini index is (2)

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2 \quad (2)$$

where p_i is the probability of an object being assigned to a specific catergory.

2) *Ada Boost:* It is a commonly used ensemble method in machine learning that aggregates multiple weak models to form a robust classifier.It trains a series of models iteratively, with each subsequent model focusing on the misclassified instances from the previous model. The algorithm allocates weights to each of its instance on the basis of difficulty in the classification task, and adjusts these weights at each iteration to prioritize the misclassified instances. The final prediction is made by aggregating the predictions of all models, with more weight given to models with higher accuracy. AdaBoost can handle both classification and regression problems and is effective in handling noisy and complex datasets. It is also less prone to overfitting compared to some other algorithms due to its focus on misclassified instances. However, AdaBoost can be sensitive to outliers and noisy data, and may not perform as well as other algorithms on datasets with high variance.The mathematical equation is given (3)

$$\alpha_t = \frac{1}{2} \ln \frac{(1 - Total\ Error)}{Total\ Error}$$

XG Boost:One of the most powerful and widely used ensemble learning algorithm that combines the strengths of decision trees and gradient boosting techniques is the Extreme Gradient Boosting algorithm It uses a series of decision trees to make predictions and iteratively improves the model by adjusting the weights of misclassified instances. XGBoost differs from traditional gradient boosting methods by incorporating a regularization component in its objective function that punishes intricate models and guards against overfitting. The parallel processing capability and ability to manage missing data makes it efficient and effective on large and complex datasets.XGBoost is frequentlty used in regression and classification problems and is commonly preferred choice for machine learning contests and real-world applications. Nevertheless, the XGBoost algorithm's effectiveness can be impacted by the selection of hyperparameters and may demand meticulous adjustment to attain peak performance.Mathematical Equation for this algorithm is (4)

$$\sum_{i=1}^{n} L(y_i , p_i) + \frac{1}{2} \lambda {O_v}^2 \quad (4)$$

where y_i is the y-axis observed value, p_i is the corresponding prediction to y values and O_v is the output value.

IV. RESULTS

Various machine learning techniques were utilized to analyze the dataset with an objective of identifying the optimal method for classifying the data. The techniques used for building the model include the creation of Decision Trees, Random Forests, Ada Boost, XGBoost and Hybrid model(RFXgB) and the accuracies obtained are shown below.

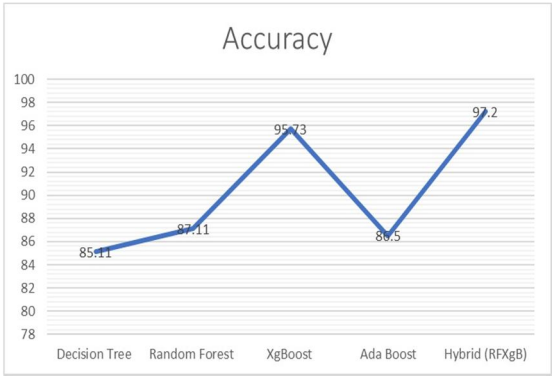


Fig 1 : Performance analysis

V. CONCLUSION

The results show the performance of four different machine learning algorithms in the task of diagnosing Polycystic Ovary Syndrome (PCOS). The algorithms evaluated are Decision Tree, Random Forest, XG Boost, and Ada Boost, and their accuracy values are presented in the table.

From the results, we can see that XG Boost achieved the highest accuracy of 95.73%, which indicates that it is the most accurate algorithm for this task among the four evaluated. Random Forest also performed well, achieving an accuracy of 87.11%, while Decision Tree and Ada Boost achieved accuracy values of 85.11% and 86.5%,and hybrid model achieved the accuracy of 97.2% respectively.

The above results suggest that machine learning algorithms have the potential to accurately diagnose PCOS based on a range of clinical and biochemical features. Hybrid Model, in particular, may be a potentially effective tool for the early detection and diagnosis of PCOS, as it achieved the highest accuracy among the evaluated algorithms.

However, it is important to note that the performance of these algorithms may vary depending on the specific dataset and features used. Additionally, more extensive research is needed to validate the above results and ensure that the algorithms are not biased or perpetuating health disparities.

REFERENCES

[1] 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.

[2] P. Mehrotra, J. Chatterjee, C. Chakraborty, B. Ghoshdastidar and S. Ghoshdastidar, "Automated screening of Polycystic Ovary Syndrome using machine learning techniques," 2011 Annual IEEE India Conference, Hyderabad, 2011, pp. 1-5.

[3] Lawrence, M.J., Eramian, M.G., Pierson, R.A. and Neufeld, E., 2007, May. It is assisted by computer to detect the polycystic ovary morphology in the ultrasound images. At Fourth Canadian Conference on Computer and Robot Vision (CRV '07) (pp. 105-112). IEEE.

[4] Cheng, J.J. and Mahalingaiah, S., 2018. Using the result reports of pelvic ultrasound classification and Data mining of ovaries. bioRxiv, p.254870.

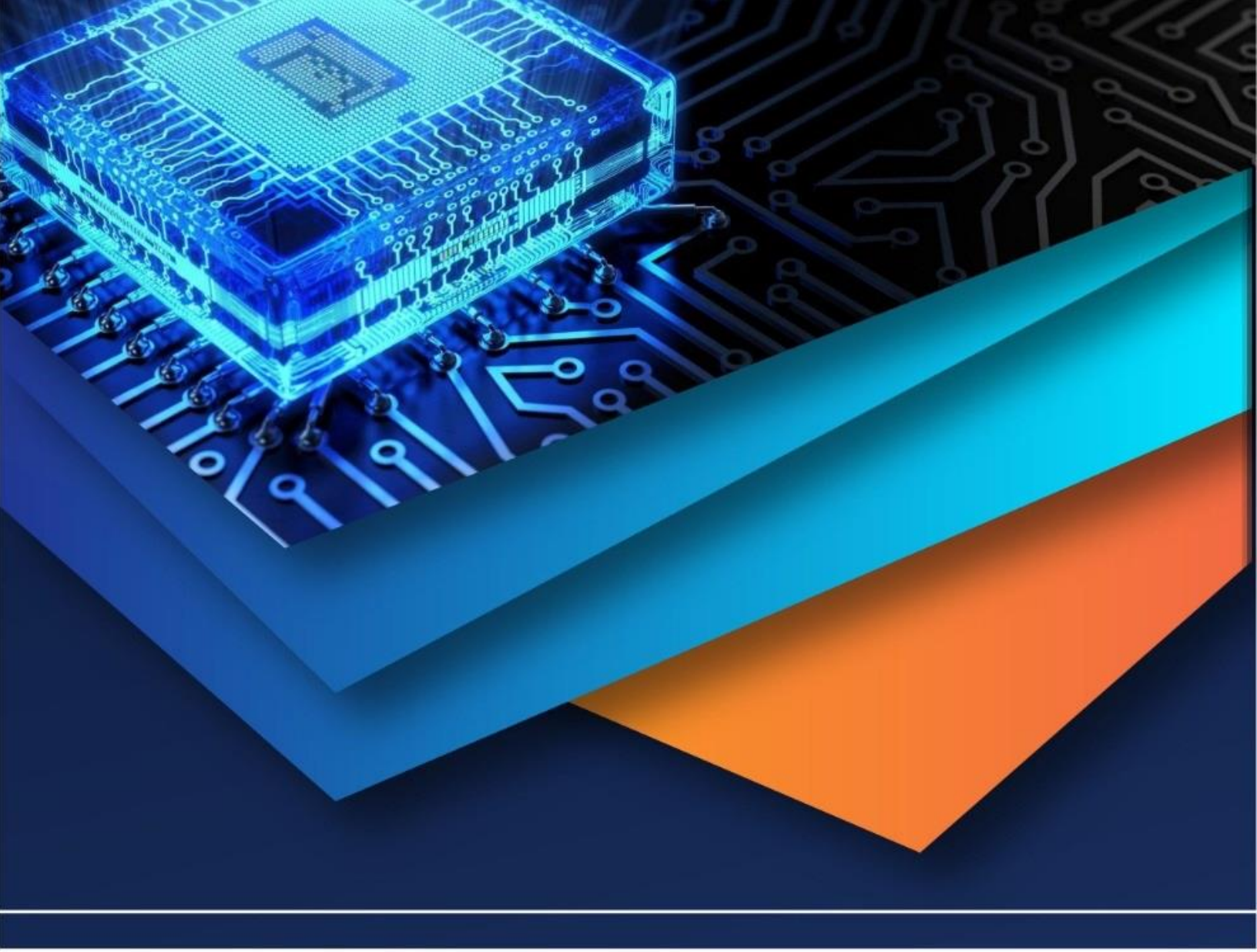
[5] Dewi, R.M. and Wisesty, U.N., 2018, March. ultrasound images using competitive neural networks. Classification of polycystic ovary based on In Journal of Physics: Conference Series (Vol. 971, No. 1, p. 012005). IOP Publishing.

[6] Sachdeva, Garima; Gainder, Shalini; Suri, Vanita; Sachdeva, Naresh1; Chopra, Seema. O Non-obese and Obese PCOS: Comparison of Clinical, Metabolic, Hormonal Parameters, and their Differential Response to Clomiphene. Indian Journal of Endocrinology and Metabolism 23(2):p 257-262, Mar–Apr 2019. | DOI: 10.4103/ijem.IJEM_637_18

[7] McCartney CR, Marshall JC. CLINICAL PRACTICE. Polycystic Ovary Syndrome. N Engl J Med. 2016 Jul 7;375(1):54-64. doi: 10.1056/NEJMcp1514916. PMID: 27406348; PMCID: PMC5301909.

[8] Norman RJ, Dewailly D, Legro RS, Hickey TE. Polycystic ovary syndrome. Lancet. 2007 Aug 25;370(9588):685-97. doi: 10.1016/S0140-6736(07)61345-2. PMID: 17720020.

[9] Essah.P.A. and Nestler, J.E., 2006. The metabolic syndrome in polycystic ovary syndrome. Journal of endocrinological investigation, 29(3), pp.270-280.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY