IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12632

Machine Learning-Based Detection and Diagnosis of Polycystic Ovary Syndrome (PCOS)

Pasula Mamatha[1] and Kadamanchi Sravani[2]

Assistant Professor, Department of Electronics and Communication Engineering, Sri Indu College of Engineering and Technology, Ibrahimpatnam, Ranga Reddy District, Telangna -501510[1]

Assistant Professor, Department of Electronics and Communication Engineering, Sri Indu College of Engineering and Technology, Ibrahimpatnam, Ranga Reddy District, Telangana -501510[2]

Abstract: Polycystic Ovary Syndrome (PCOS) is a condition that affects women during their reproductive years. This project aims to reduce the risk of serious health complications by enabling early detection of PCOS through advanced machine learning techniques. Using a dataset from Kaggle that includes both clinical and physical attributes of women, the project focuses on predicting PCOS effectively. Additional features integrated into the system include a menstrual cycle tracker, customized diet and yoga plans, PCOS detection via ultrasound imagery, and access to virtual doctor consultations. To support this, three distinct machine learning models have been developed: **PCOS Model 1**, which achieved 97% accuracy using the XGBoost algorithm; **PCOS Model 2**, with 92% accuracy using Random Forest; and the **Image PCOS Model**, which attained 96% accuracy using a Convolutional Neural Network (CNN). These models significantly enhance early diagnosis efforts and contribute to creating a holistic, user-friendly platform for managing women's health.

Keywords: Consultation, detection, hormonal, PCOS, XGBoost, ovary, menstrual etc.

I. INTRODUCTION

PCOS a multifaceted endocrine disorder occurring in women during their reproductive years, poses challenges in both diagnosis and treatment. It is addressed by various symptoms such excessive hair growth, irregular periods, and weight gain. One of the primary hurdles in effectively addressing PCOS lies in its diagnostic ambiguity. The varied and often overlapping symptoms among individuals, coupled with the absence of a definitive singular diagnostic criterion, render the identification and classification of this syndrome a daunting task for healthcare practitioners. The intricacies in diagnosis stem from the absence of a standardized definition for PCOS, leading to divergent diagnostic criteria put forth by various expert organizations like the National Institute of Health Consensus Statement, ESHRE/ASRM, and the Androgen Excess and PCOS Society. This ambiguity in diagnosis contributes to a notable proportion of cases going unrecognized, with estimates indicating that as much as 70% of the affected women remain undiagnosed. Moreover, the current diagnostic methodologies, relying heavily on clinical tests and imaging procedures, often necessitate extensive examinations, leading to increased healthcare costs, patient discomfort, and delays in timely intervention.

The advancements in fields of Machine Learning and Artificial Intelligence provides us an exposure to deal with problems of PCOS [7]. It also grants us the capability to process huge sets of clinical data and patterns, and learn from these datasets without explicit programming. Its application in healthcare, specifically in the realm of disease diagnosis. It aims to leverage machine learning's capabilities in tackling PCOS diagnosis by crafting an all- encompassing diagnostic framework employing advanced algorithms facilitated by machine learning techniques. The study extensively utilizes a robust dataset obtained from Kaggle to build and assess the efficacy of this model. Leveraging seven different classifiers, the research meticulously compares their accuracy and juxtaposes the results with established literature employing similar datasets. The importance of this study rests in its capacity to improve and simplify the diagnostic procedure for PCOS. [9]. By acquiring the power and capability of machine learning algorithms this study aims to provide a more efficient, accurate, and accessible diagnostic tool for healthcare practitioners. The goal is to facilitate early identification, personalized intervention, and improved management strategies for women affected by PCOS. This innovative approach not only holds promise for revolutionizing PCOS diagnosis but also in the underscore's transformative role of the technology in healthcare.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12632

II. PROPOSED SYSTEM

In figure 1, we outline the step-by-step process undertaken to construct and train our machine learning model. The methodology encompasses key stages from the initial acquisition of data to the evaluation of model accuracy.

1. Labelled Dataset

Our study began with the acquisition of a well- structured dataset containing labeled examples, wherein each mentioned data point was linked with the predefined variables. The selection of this dataset was crucial to ensure representation and diversity, mirroring the real- world scenarios our model is intended to address.

2. Data Pre-processing

To prepare the dataset for model training, comprehensive data preprocessing was undertaken. This involved handling missing data through imputation or removal, cleaning the dataset by addressing outliers and errors, and standardizing or normalizing numerical features. Additionally, categorical variables were encoded to transform them into a numerical format suitable for machine learning algorithms

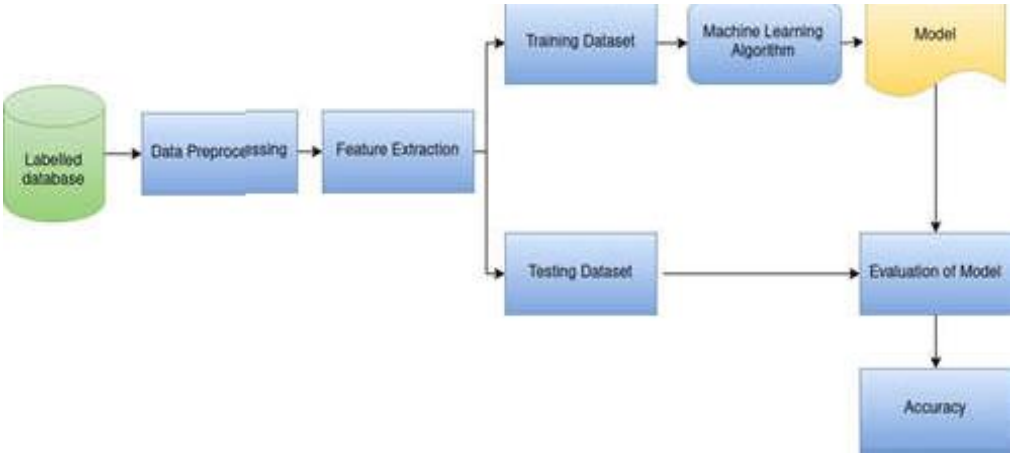


Figure 1. The methodological architecture for creation of Machine Learning Model

3. Feature Extraction

The identification of relevant features played a pivotal role in enhancing the model's predictive capabilities. Feature extraction required a meticulous examination to choose features that made substantial contributions to the model's overall effectiveness.

4. Division of Dataset into Training and Testing Sets

The labeled dataset was split into training and testing sets. Training data facilitated model learning, while testing data evaluated its generalization and performance on unseen instances.

5. Choosing of Machine Learning Algorithm

The selection of a well-suited machine learning algorithm was based on the nature of the problem at hand. The chosen algorithm underwent iterative training, with adjustments made to its parameters to optimize performance. The training process involved refining the model through continuous evaluation on the training dataset.

6. Model Training

The actual training of the machine learning model took place using the preprocessed and feature-engineered dataset. The model learned patterns from the training set, and adjustments were made to its parameters to enhance its predictive accuracy. This iterative process aimed to strike a balance between under fitting and overfitting.

7. Evaluation of Model

To gauge the model's performance, it was evaluated using the testing dataset. It was tested using the evaluation dataset. Metrics such as accuracy, precision, recall, and F1 score were utilized to gain insight into the model's capabilities and limitations. Cross-validation methods were also utilized to evaluate its consistency across various data subsets.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12632

8. Accuracy

While accuracy served as a primary metric for measuring the correctness of predictions, we acknowledged its limitations, especially in the context of unbalanced datasets. Therefore, additional metrics were considered to provide a more nuanced assessment of the model's effectiveness.

Fig 2 begins with a carefully curated dataset; we utilize the CS-PCOS feature engineering method to enhance predictive accuracy. An exploratory data analysis informs subsequent steps, including training/testing set division. The trained model predicts PCOS instances, and evaluation metrics such as accuracy, precision, recall, and F1 score assess its performance. This systematic approach ensures a comprehensive understanding of PCOS prediction, from data input to syndrome detection.

Table 1 is the dataset that comprised of 26 features, each providing valuable insights into individuals' health and lifestyle attributes, with a specific focus on factors relevant to Polycystic Ovary Syndrome (PCOS) detection. The attributes encompass diverse data types, including integers and floats, catering to various aspects of health measurement and personal characteristics.

For instance, demographic details such as age, weight, and height are represented as integers and floats, while categorical information like blood group and pregnancy status is expressed through integer values. With 1082 non- null entries for each attribute, the dataset ensures a robust and consistent dataset for research purposes.

This comprehensive collection includes reproductive factors, body measurements, lifestyle habits, and blood pressure readings, culminating in an encompassing exploration of potential correlations and risk factors associated with PCOS. Researchers can leverage this dataset to conduct thorough analyses, contributing valuable insights to the understanding and diagnosis of PCOS in clinical and research settings.

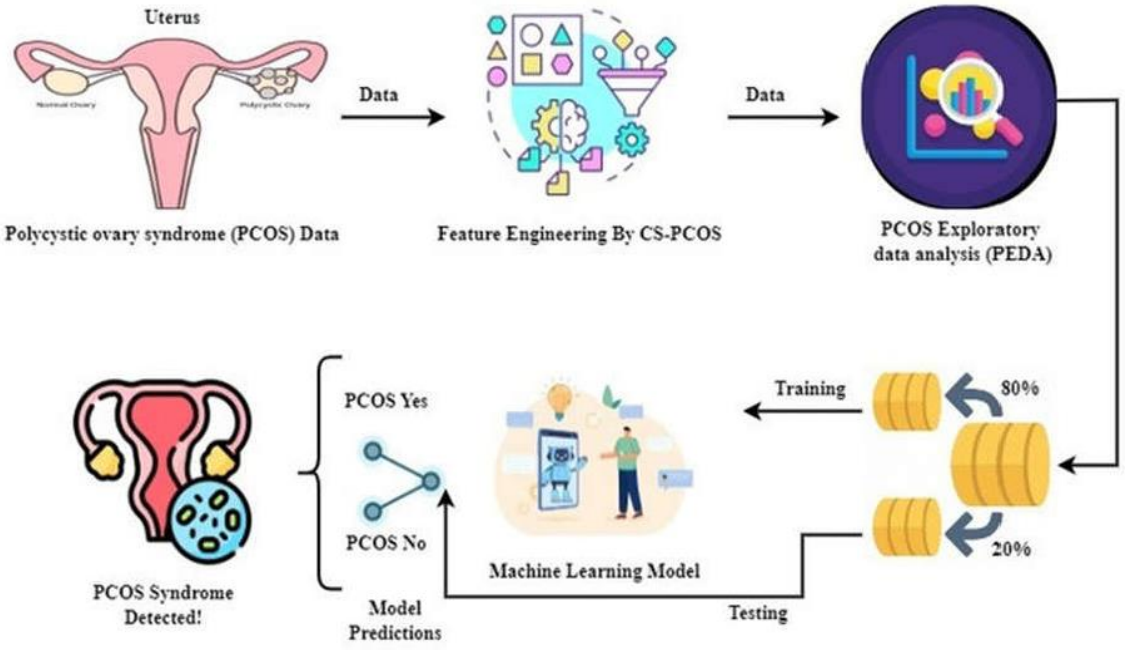


Figure 2. The methodological architectural analysis of the proposed research study in predicting the PCOS syndrome

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12632

Sr. No.	Feature	Not- Null Count	Data Type	Sr. No.	Feature	Not- Null Count	Data Type
1	Age (yrs)	1082	int64	14	Hip(inch)	1082	int64
2	Weight (Kg)	1082	float64	15	Waist(inch)	1082	int64
3	Height(Cm)	1082	float64	16	Waist-Hip Ratio	1082	float64
4	BMI	1082	float64	17	Weight gain(Y/N)	1082	int64
5	Blood Group	1082	int64	18	hair growth(Y/N)	1082	int64
6	Pulse rate(bpm)	1082	int64	19	Skin darkening (Y/N)	1082	int64
7	RR (breaths/min)	1082	int64	20	Hair loss(Y/N)	1082	int64
8	Hb(g/dl)	1082	float64	21	Pimples(Y/N)	1082	int64
9	Cycle(R/I)	1082	int64	22	Fast food (Y/N)	1082	int64
10	Cycle length(days)	1082	int64	23	Reg Exercise(Y/N)	1082	int64
11	Marraige Status (Yrs)	1082	float64	24	BP _Systolic (mmHg)	1082	int64
12	Pregnant(Y/N)	1082	int64	25	BP _Diastolic (mmHg)	1082	int64
13	No. of abortions	1082	int64	26	PCOS (Y/N)	1082	int64

Table 1. The PCOS dataset descriptive feature analysis

III. EXPERIMENT AND RESULT

D analysis of feature distribution analysis of class.

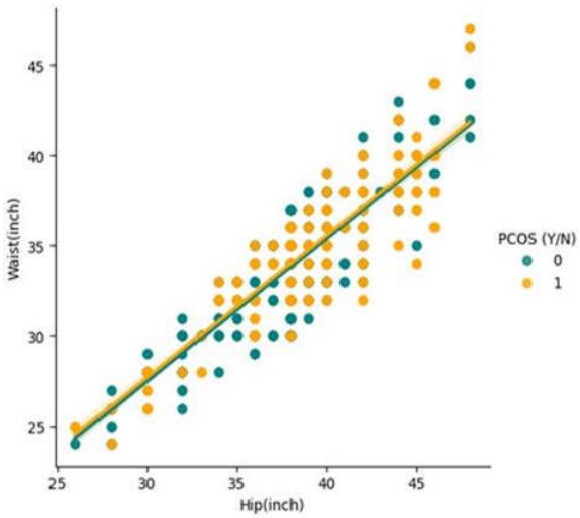


Figure 3. The implot is drawn on feature Hip: Waist

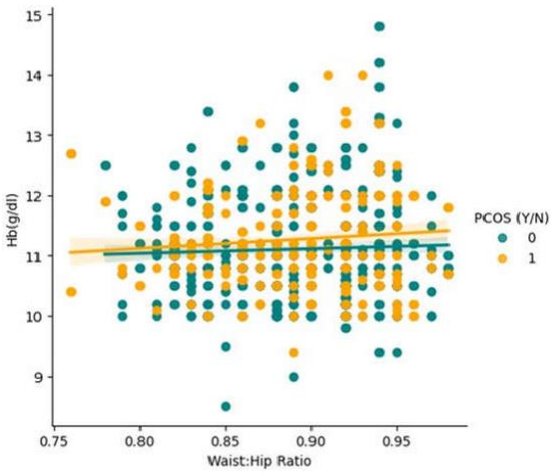


Figure 4. The implot is drawn on feature Waist: Hip: Hb

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12632

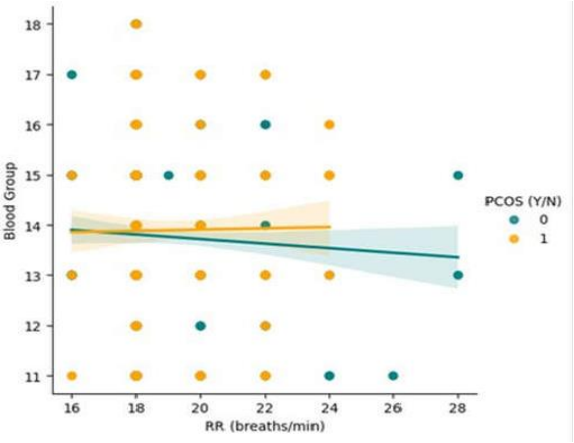


Fig.5 The implot is drawn on feature Blood Group: RR

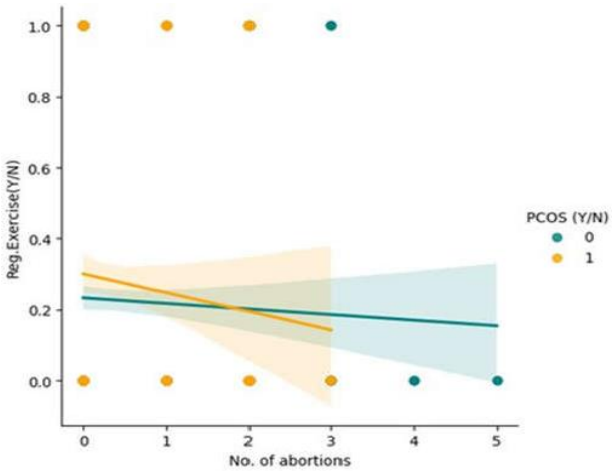


Fig 6.The implot is drawn on feature No of abortion: Reg Exercise

Employment of different machine learning Algorithms

Below describes the brief explanation of each classification algorithm we have used for the detection of Polycystic Ovary Syndrome (PCOS):

1. **Support Vector Machine (SVM):**

SVM is a powerful algorithm that finds the best possible separation between different classes. It aims to create a hyperplane that maximizes the margin between data points of different classes.

2. **Logistic Regression:**

Logistic Regression, a straightforward yet powerful technique for binary classification. It predicts the probability of instances belonging to a specific class and makes predictions using the logistic function.

3. **CatBoost:**

CatBoost is a gradient boosting method specifically crafted to handle categorical features effectively. Renowned for its capability in efficiently managing categorical data, it frequently necessitates minimal hyperparameter adjustment.

4. **Decision Tree:**

Decision Trees use a tree model to make decisions based on features, with each node representing a feature and each branch a decision. They are known for being interpretable and easy to understand.

5. **XGBoost:**

XGBoost, or Extreme Gradient Boosting, is a potent ensemble learning algorithm ideal for classification and regression tasks.

6.K-Nearest Neighbors (KNN):

KNN is a straightforward and intuitive algorithm that classifies instances by considering the majority class of their nearest neighbors. It assesses similarity using distance measures like the Euclidean distance.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12632

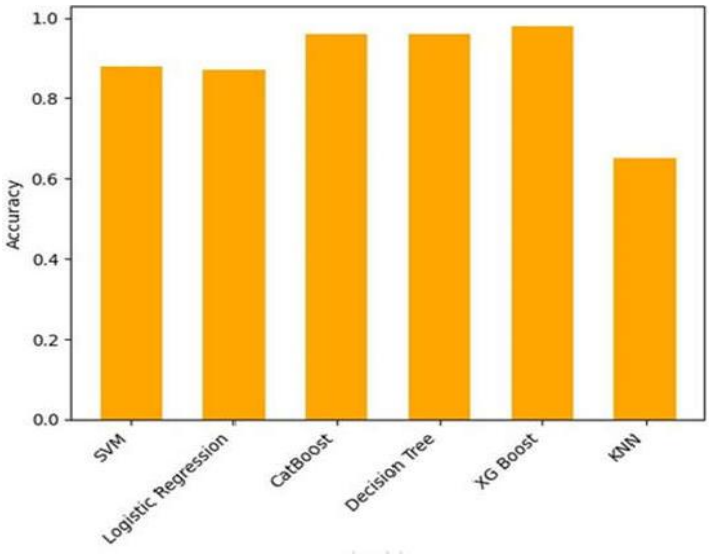
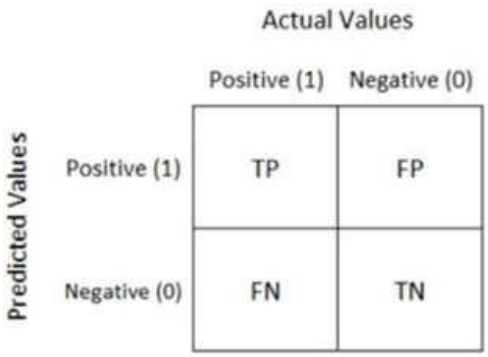


Figure 7. Comparison of different machine learning algorithms

So, figure 7 illustrate the classification accuracy achieved by different algorithms for the detection of Polycystic Ovary Syndrome (PCOS). The results showcase the efficacy of each algorithm in capturing patterns within the dataset. Specifically, Support Vector Machine (SVM) demonstrated a commendable accuracy of 85%, closely followed by Logistic Regression at 84%. Notably, CatBoost outperformed others with an impressive accuracy of 94%, while Decision Tree and XGBoost exhibited high accuracies of 95% and 97%, respectively. However, K- Nearest Neighbors (KNN) showed a comparatively lower accuracy of 68%. These findings offer valuable insights into the performance variations among the employed algorithms, aiding in the selection of optimal models for PCOS detection.

Evaluation of Model

Multiple methods exist for assessing the efficacy of machine learning models. We assessed model performance using the following performance metrics. A confusion matrix, which summarizes predictions in a tabular format, is utilized to gauge model performance. It comprises a combination of predicted and actual values, providing insight into the model's accuracy. where, True Positive (TP) occurs when the model correctly identifies individuals with PCOS, while True Negative (TN) refers to instances where the model accurately identifies individuals without PCOS. False Positive (FP), or Type I Error, happens when the model incorrectly predicts PCOS in individuals who do not have it. False Negative (FN), or Type II Error, arises when the model fails to predict PCOS in individuals who have it.



In the context of our research, accuracy returns the value of how many cases related to PCOS we correctly labelled out all the cases.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Precision refers to how many of people labelled as PCOS have PCOS.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12632

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

Recall and Sensitivity refers to the value of how many women have PCOS, how many of them are correctly predicted?

$$\text{Sensitivity} = \frac{TP}{(TP + FN)}$$

F1 Score refers to the average of precision and recall

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Specificity})}{\text{Precision} + \text{Specificity}}$$

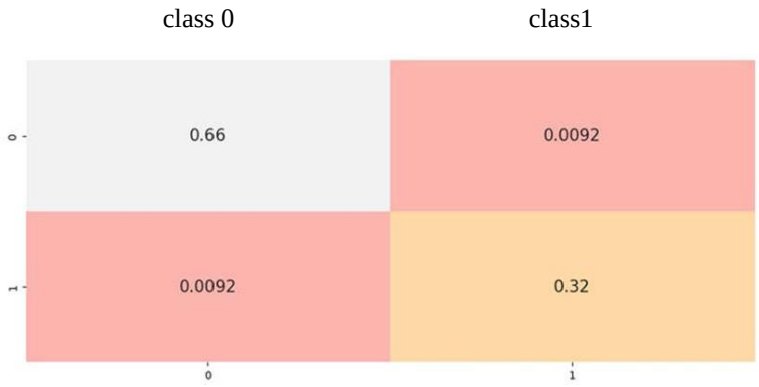


Figure 8. Performance measure for Machine Learning Model

```
Confusion Matrix:
[[144   2]
 [  2  69]]
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       146
           1       0.97      0.97      0.97        71

    accuracy                           0.98       217
   macro avg       0.98      0.98      0.98       217
weighted avg       0.98      0.98      0.98       217
```

Figure 9. Comparison Analysis of accuracy of each class

Figure 8, indicates the performance metrics of our advanced machine learning model, specifically employing Random Forest, for the further classification of Polycystic Ovary Syndrome (PCOS) types. The model achieved an overall accuracy of 79%, indicating its effectiveness in correctly classifying instances. Additionally, we evaluated the model's performance across different classes. For Class 0 (PCOS present), precision, sensitivity, and F1 score metrics were calculated. The same set of metrics was applied to Class 1 (PCOS Absent), These detailed performance measures offer a comprehensive understanding of the model's efficacy in distinguishing between the various PCOS subtypes.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12632

In figure 9, we illustrate a comprehensive overview of the base machine learning model's performance metrics. This model was specifically employed for the detailed classification of Polycystic Ovary Syndrome (PCOS) detection. The diagram includes crucial metrics such as accuracy, precision, sensitivity, and F1 score, providing a thorough assessment of the model's effectiveness across all four classes. These classes, denoted as Class 0 (PCOS present), Class 1 (PCOS Absent), undergo individual evaluation for their classification performance. This granular analysis facilitates a nuanced understanding of the model's capabilities in accurately identifying the PCOS.

IV. CONCLUSION

Polycystic Ovary Syndrome (PCOS) is a problem of concern for women, worldwide linked to conditions like preterm abortions, infertility, and anovulation. Its diverse symptoms, including irregular menstrual cycles, obesity, and hirsutism, make timely diagnosis challenging. The associated clinical tests and ovarian scanning procedures are not only burdensome but also time-consuming and expensive for patients. To address this issue in the early detection and improved health outcomes for PCOS patients by analyzing input text and sonography reports. We also enable timely diagnosis and offer a range of features, including period tracking and a BMI calculator, to empower individuals to manage their health comprehensively. In addition, we provide doctor consultations for severe issues and deliver customized diet plans, yoga, and exercise recommendations. PCOS classification involves employing different machine learning methods like Naïve Bayes, logistic regression, K-Nearest Neighbor (KNN), Classification and Regression Trees (CART), Random Forest Classifier, and Support Vector Machine (SVM) within the Jupyter Python IDE. Findings indicate that XGBoost emerges as the most appropriate and precise method for PCOS prediction, achieving an accuracy rate of 98.16%.

REFERENCES

[1]. S. Dhinakaran, C. Thangavel, S. S and H. V S, "PCOS Perception analysis prediction using Machine learning algorithms," 2022 IEEE 7th International Conference on Recent Advances and Innovations In Engineering (ICRAIE), MANGALORE, India, 2022, PP:260-265 DOI:10.1109/ICRAIE56454.2022.10054279.

[2]. S. Nasim, M. S. Almutairi, K. Munir, A. Raza and F. Younas, "A Novel Approach of Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics," in IEEE Access, vol. 10, pp: 97610-97624, 2022, doi: 10.1109/ACCESS.2022.3205587.

[3]. P. Chauhan, P. Patil, N. Rane, P. Raundale and H. Kanakia, "Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS," 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 2021, pp. 1-7, doi: 10.1109/ICCICT50803. 2021.9510128.

[4]. V. Srinithi and R. Rekha, "Machine learning for diagnosis of polycystic ovarian syndrome (PCOS/PCOD)," 2023 International Conference on Intelligent Systems for Communication, India, 2023, pp. 19-24, doi: 10.1109/ICISCoIS56541.2023.10100490.

[5]. A. Denny, A. Raj, A. Ashok, C. M. Ram and R. George, "i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 673-678, doi: 10.1109/TENCON.2019.8929674.

[6]. D. Hdaib, N. Almajali, H. Alquran, W. A. Mustafa, W. Al-Azzawi and A. Alkhayyat, "Detection of Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms," 2022 5th International Conference on Engineering Technology and its Applications (IICETA), Al-Najaf, Iraq, 2022, pp. 532-536, doi: 10.1109/IICETA54559.2022.9888677.

[7]. Y. A. Abu Adla, D. G. Raydan, M. -Z. J. Charaf, R. A. Saad, J. Nasreddine and M. O. Diab, "Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques," 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME), Werdanyeh, Lebanon, 2021, pp. 208-212, doi: 10.1109/ICABME53305.2021.9604905.

[8]. P. B and R. Khilar, "Classification of PCOS Using Machine Learning Algorithms Based on Ultrasound Images of Ovaries," 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICONSTEM56934.2023.10142359.