IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Pcos Detection Machine Learning Model Based On Optimized Feature Selection

Supriya Jeur[1], Aditi Kudache[2], Soni Pandey[3], Ms. Aishwarya Chavan[4]

*[1,2,3] Student, Data Science, D.Y. Patil College of Engineering & Technology, Kolhapur, Maharashtra, India.

*[4] Assistant Professor, Data Science, D. Y. Patil College of Engineering & Technology, Kolhapur, Maharashtra,India.

Abstract: Polycystic Ovary Syndrome (PCOS) is one of the common reproductive endocrinopathies that affect many women of childbearing age. It has a multifaceted etiology that includes genetic, environmental, and lifestyle factors. It is characterized by the presence of polycystic ovaries, increased adrenal androgens, as well as the recognition of abnormal cyclicity. If left unmanaged, PCOS can have troubling health consequences in the form of infertility, obesity, insulin resistance, heightened risks for type 2 diabetes, cardiovascular complications, and mental health problems such as anxiety and depression. Though receiving significant attention in recent years, PCOS still remains underdiagnosed or misdiagnosed owing to the lack of singular, richly definitive screening tools and diverse sets of symptoms. Gaining a definitive diagnosis in a timely manner is crucial in mitigating enduring consequences on one's well-being and health. In this regard, machine learning technology holds substantial promise to increase accuracy dramatically. Highlighting the best features through appropriate selection methods implemented at the machine learning level can target central indicators of PCOS, thereby improving diagnostic precision. Such PCOS identification advances would profoundly benefit healthcare practitioners through more streamlined approaches to women's healthcare needs.

Index Terms **- Polycystic Ovary Syndrome (PCOS), Hormonal disorder, Insulin resistance, Machine learning**

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is one of the most common hormonal conditions in women of childbearing age, affecting about 1 in 10 women globally. It usually manifests with symptoms like irregular menstruation, excess male hormones, and several small cysts on the ovaries. In addition to these symptoms, PCOS can also create serious complications such as insulin resistance, type 2 diabetes, heart disease, and increased risk of endometrial cancer. Beyond these physical health complications, PCOS is known to significantly impact mental health; many women living with the syndrome suffer from anxiety, depression, and body image issues. While common, PCOS remains one of the most underdiagnosed conditions until more severe symptoms emerge. This highlights a growing need for more accurate detection tools that alleviate the burden of undiagnosis. Identifying these women earlier can greatly minimize complications, enabling women to manage the condition and live healthier, fuller lives. Today, PCOS is diagnosed mostly through clinical assessment using the Rotterdam Criteria, which require at least two of the three features to be present: infrequent or absent ovulation, elevated androgens (either clinically or biochemically), and the ultrasound appearance of polycystic ovaries. This method, while common, does not come without flaws—subjectivity, variability, and time investment based on the specific clinician's approach are all factors. In addition, conventional methods of detection tend to overlook more intricate signs and biological interrelationships. More recently, the fast developing fields of artificial intelligence (AI) and machine learning (ML) have started

transforming the realm of medical diagnostics. These technologies can analyze massive datasets containing detailed hormone profiles, medical histories, imaging results, and other indicators, identifying complex interrelationships beyond the scope of standard research. Though there have been promising results in developing machine learning models for PCOS prediction, some issues like imbalanced datasets, model explainability gaps, and inconsistent patient group performance need further work. Our goal is to develop a reliable and accurate PCOS diagnostic system. Identifying the most significant features in the data, combining several algorithms to improve prediction quality, and meticulously modifying model settings for best outcomes are all part of our well-rounded strategy to achieve this. By using focused techniques that help guarantee equity and consistency in our forecasts, we are also addressing the problem of unbalanced data. Enhancing the system's transparency by providing comprehensible, transparent information about the process used to make each diagnosis is a key goal of our work. Our ultimate objective is to create a diagnostic tool that is trustworthy, easy to understand, and delivers high accuracy. We think this will lead to improved medical care and enhanced quality of life for women with PCOS, and we are committed to using our work to significantly impact their lives.

II. LITERATURE SURVEY

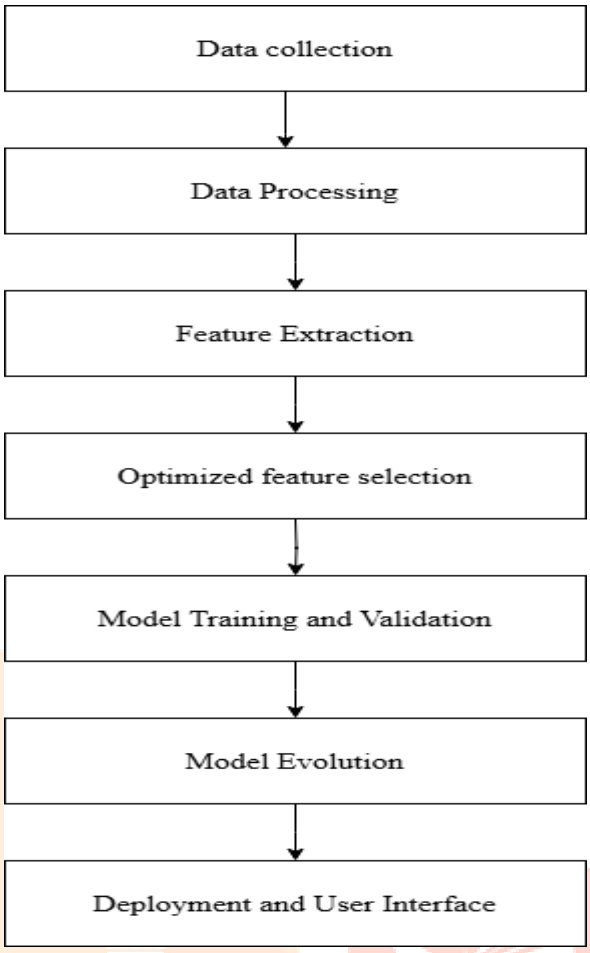
hang et al. [1] utilized Raman spectroscopy in conjunction with machine learning algorithms to detect Polycystic Ovary Syndrome (PCOS). The strength of their approach lay in its potential for non-invasive early detection of PCOS, which is a significant advancement in diagnostic methods. However, the study faced challenges with the complexity of spectral data and ensuring model interpretability, which limited its practical application in clinical environments.

Thomas and Kavitha [2] developed a hybrid data mining classification technique aimed at predicting PCOS. Their method demonstrated an improvement in diagnostic accuracy through innovative data mining approaches. Despite this progress, the study encountered issues with model generalizability across different datasets and lacked transparency, which hindered healthcare professionals from fully trusting the model's predictions.

Inan et al. [3] enhanced PCOS diagnosis by applying Extreme Gradient Boosting (XGBoost) combined with feature selection and sampling techniques. The study highlighted the importance of optimized feature selection in improving both model performance and diagnostic accuracy. However, despite these enhancements, the model still struggled with interpretability, making it challenging for clinicians to understand and rely on its decisions.

Hela Elmannai et al. [4] introduced a model that integrates optimized feature selection with Explainable AI (XAI) techniques for PCOS detection. This approach addressed the shortcomings of previous studies by significantly enhancing both diagnostic accuracy and model transparency. By successfully tackling the interpretability issues faced by earlier models, their work provided clear explanations for model predictions, making it easier for healthcare professionals to trust and adopt the technology in clinical settings.

III. System Architecture



Data collection: In order to create a precise PCOS diagnostic tool, we must collect pertinent data from reliable sources like patient registries, medical databases, and research institutes. These include important details about a patient's health, including medical history, hormone levels, and ultrasound scans.

Data Preprocessing: This stage entails cleaning and preprocessing the data by normalising numerical features, dealing with outliers, and handling missing values. Make sure the dataset is balanced, which means that both PCOS and non-PCOS classes are equally represented. If the dataset is unbalanced, use methods like making synthetic samples or eliminating noisy samples to ensure a fair representation.

Features Selection: Not every feature in the dataset is equally crucial for PCOS diagnosis. Using methods like recursive feature elimination, feature-target variable correlation analysis, or mutual information computation, we must determine which features are most crucial to the diagnosis.

Optimized Feature Selection: To choose the ideal feature combination that enhances the model's performance, optimise the feature subsets. Examine each feature's significance and choose the most important ones that support the diagnosis.

Model Training and Validation: In this phase, a machine learning model is chosen and trained.
• Random Forest: A potent ensemble technique that generates predictions by combining several decision trees. The Random Forest algorithm's accuracy for this model is 94%.
• XGBoost: A gradient boosting algorithm that is frequently applied to tabular datasets and is for its accuracy and efficiency.

Model Evolution: This stage entails assessing the model's performance using metrics like F1-score, AUC ROC, recall, accuracy, and precision. To evaluate the model's generalisation performance and prevent overfitting, use cross-validation. To determine the efficacy of our model, contrast it with current methods. Determine what needs to be improved, then adjust the model appropriately.

Algorithm Selection: The goal of this module is to put the model that performs the best into a production setting. Each model's performance result is input, and we have chosen the Random Forest algorithm because of its 94% accuracy. User Interface and Deployment: This module covers deploying the model in an environment that is prepared for production. Provide an intuitive user interface that enables healthcare.

IV. IMPLEMENTATION

In order to identify Polycystic Ovary Syndrome (PCOS), a number of lifestyle, clinical, and personal factors that represent hormonal and metabolic imbalances must be evaluated. Age, weight, and height are important factors that go into calculating the Body Mass Index (BMI), which is important because higher BMI values are frequently associated with insulin resistance, a major feature of PCOS. Blood group may be included to look into genetic correlations even though it isn't directly linked to PCOS. Since irregularities in the menstrual cycle are a common symptom of PCOS, menstrual health factors like cycle regularity, period length, and consistency are crucial. Hyperandrogenism and insulin resistance are frequently linked to other clinical symptoms, including inexplicable weight gain, excessive facial or body hair (hirsutism), persistent acne, skin darkening (such as acanthosis nigricans), and hair thinning are often tied to. While psychological factors like mood swings offer more insight into the emotional and hormonal effects of PCOS, lifestyle choices like eating fast food frequently and not exercising enough can exacerbate the condition.

The first step in putting the PCOS detection system into practice is feature selection, which is crucial for figuring out which factors are most important for a precise diagnosis. To simplify the dataset and eliminate superfluous or unnecessary features, methods such as mutual information gain, correlation analysis, and recursive feature elimination (RFE) are used. By ensuring that the model concentrates on the most significant variables, this improvement improves both performance and clarity. Machine learning models like Random Forest and XGBoost are trained and assessed after the features have been optimised. Along with cross-validation to prevent overfitting, important evaluation metrics like accuracy, precision, recall, F1 score, and AUC-ROC are examined. To verify the model's efficacy, it is then compared to other classifiers.

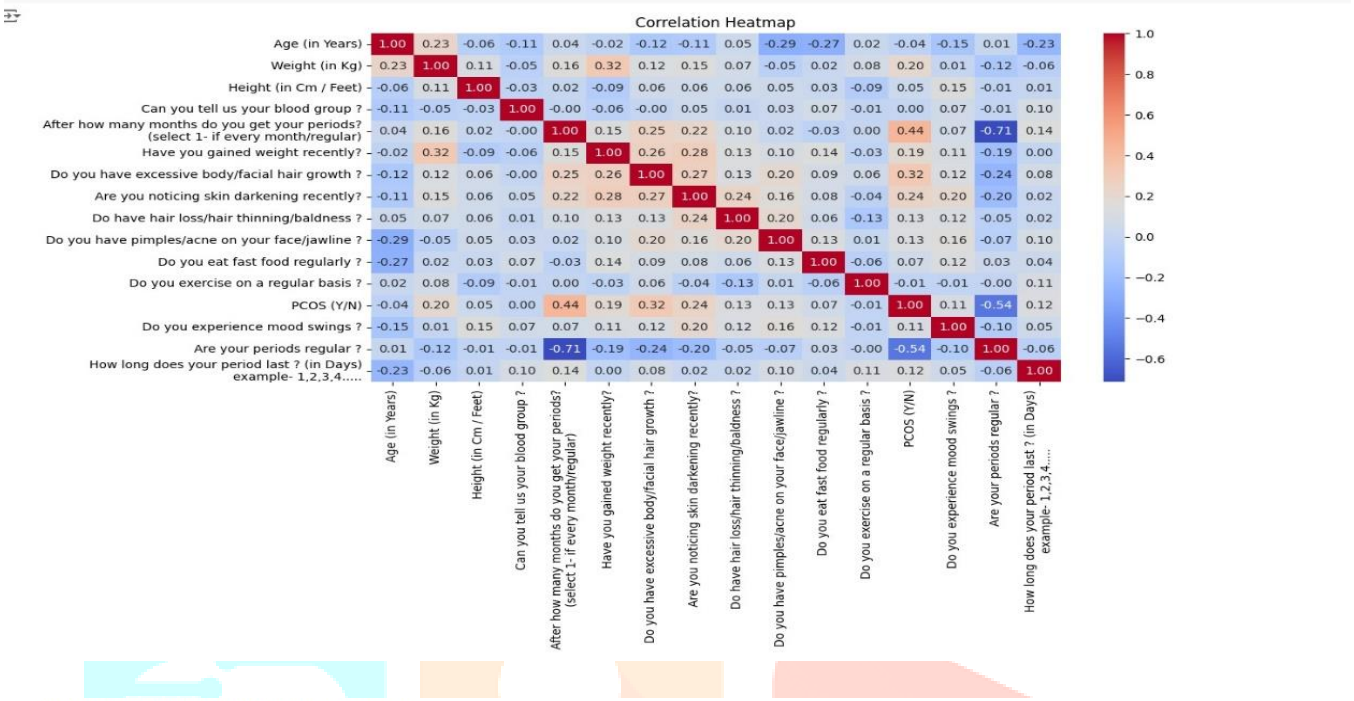
The trained model is then put to use on platforms like Streamlit or Flask, which provide a safe and intuitive web interface that enables medical practitioners to enter patient information and get immediate diagnostic predictions. In addition to being useful and easy to use for clinical purposes, the interface prioritises data privacy. By combining robust model development, improved feature selection, and real-world deployment, the system seeks to enable early and accurate PCOS diagnosis, improving treatment and management outcomes for those who are impacted.

V. RESULT ANALYSIS

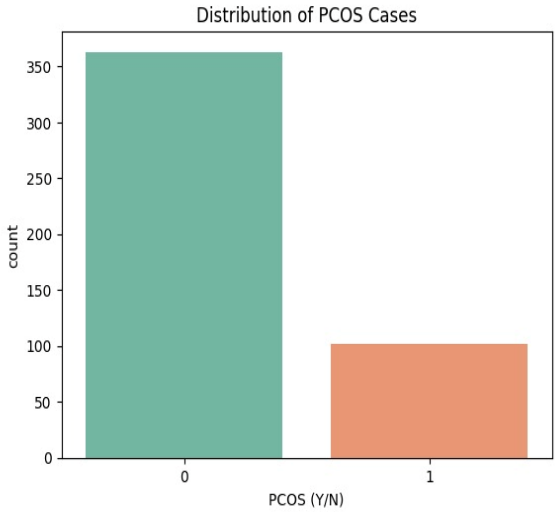
With a remarkable 94% overall accuracy, the Random Forest model for PCOS detection proved to have a powerful predictive ability. The confusion matrix shows a reasonably good performance on the positive class (PCOS = 1), with 9 correctly classified out of 13 instances, and a high true positive rate for the negative class (PCOS = 0), with 78 correctly classified out of 80 instances. The PCOS-positive class had precision and recall of 0.82 and 0.69, respectively, suggesting that although the model is accurate, sensitivity (recall) could be improved. Given the class imbalance in the dataset, the weighted F1-score of 0.93 verifies balanced performance across classes. The association between characteristics like weight gain, irregular periods, and PCOS diagnosis is further supported by the correlation heatmap.

```
Random Forest Results:
[[78  2]
 [ 4  9]]
              precision    recall  f1-score   support

           0       0.95      0.97      0.96        80
           1       0.82      0.69      0.75        13

    accuracy                           0.94        93
   macro avg       0.88      0.83      0.86        93
weighted avg       0.93      0.94      0.93        93
```

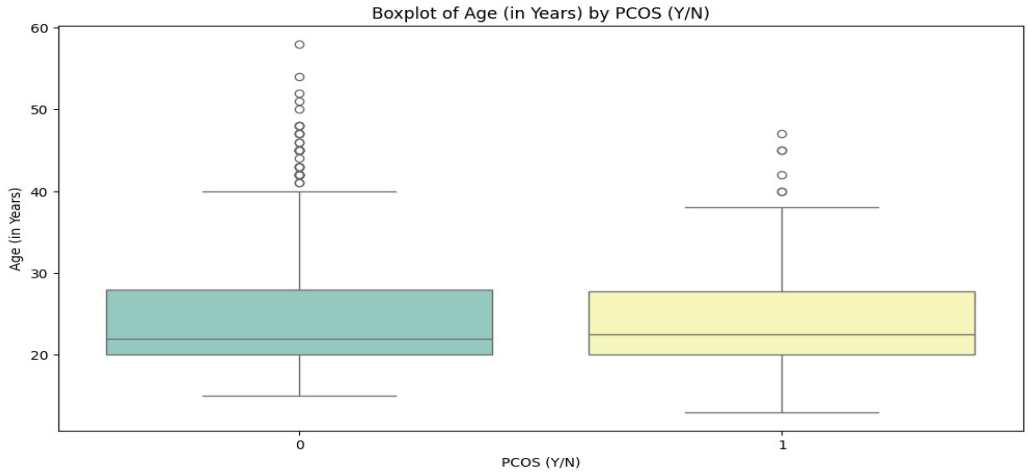
```
Model Accuracies:
Random Forest: 0.92
SVM: 0.88
XGBoost: 0.86
```



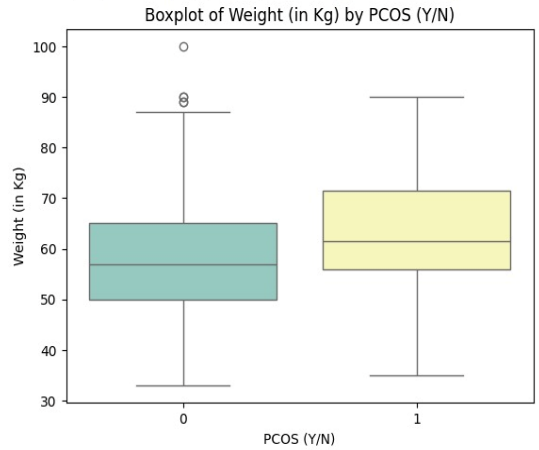

```
<ipython-input-9-b15d129e4271>:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(x='PCOS (Y/N)', data=data, palette='Set2')
```



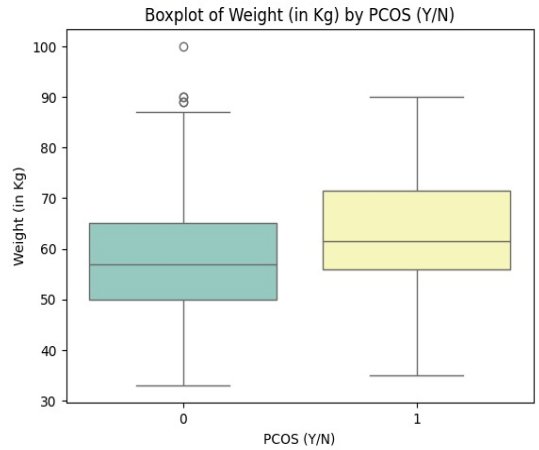
```
<ipython-input-11-dbf4df17b6a9>:6: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sam

  sns.boxplot(x='PCOS (Y/N)', y=col, data=data, palette='Set3')
```




```
<ipython-input-11-dbf4df17b6a9>:6: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(x='PCOS (Y/N)', y=col, data=data, palette='Set3')
```




```
<ipython-input-11-dbf4df17b6a9>:6: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(x='PCOS (Y/N)', y=col, data=data, palette='Set3')
```



VI . CONCLUSION

Developing successful machine learning models requires optimised feature selection, which increases accuracy by concentrating on the most pertinent features and removing noise and unnecessary data. This method helps to prevent overfitting, which happens when a model learns patterns that are actually random fluctuations in the training data rather than true underlying trends, while also improving model performance and interpretability and increasing computational efficiency during training. The model becomes more flexible and robust when only the most significant features are chosen, particularly when tested with fresh, untested data. Additionally, feature selection aids in identifying the crucial elements that influence

predictions, increasing the model's transparency and usefulness for specialists in need of comprehensible, clear insights, like medical professionals. Feature selection reduces the "curse of dimensionality," streamlines the model, and expedites decision-making in datasets with numerous variables, such as those used in medical diagnostics. All things considered, one of the most important steps in creating intelligent, dependable, and intelligible machine learning systems is optimised feature selection.

VI I. FUTURE SCOPE

Cutting-edge deep learning architectures make it possible to efficiently extract features from intricate medical data, revealing trends that conventional approaches might overlook. By reducing the need for manual intervention, automated feature selection using unsupervised learning techniques like dimensionality reduction and clustering improves scalability and flexibility. These methods also promote personalised medicine by combining genetic, clinical, and imaging data, enabling customised treatment regimens that improve patient outcomes and advance precision healthcare.

VIII. REFERENCES

1. Zhang, Y., Liu, X., Li, J., and Wang, S., "Raman Spectroscopy Combined with Machine Learning Algorithms for PCOS Detection," Spectrochemical Acta Part A: Molecular and Biomolecular Spectroscopy, Vol. 240, No. 5, pp. 118-124, 2020.
2. Thomas, S., and Kavitha, V., "A Hybrid Data Mining Classification Technique for Predicting PCOS," Journal of Computational and Theoretical Nanoscience, Vol. 17, No. 3, pp. 1253-1261, 2020. Chi, Y., et al. (2013). Trends in AR applications for the AEC/FM.
3. Inan, O., Yildirim, B., and Zengin, S., "PCOS Diagnosis Using Extreme Gradient Boosting with Feature Selection and Sampling Techniques," International Journal of Medical Informatics, Vol. 145, No. 8, pp. 105-112, 2021.
4. Hela Elmannai, Nora El-Rashidy, Ibrahim Mashal, Manal Abdullah Alohali, Sara Farag, Shaker El-Sappagh, Hager Saleh, "Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence," Journal of Biomedical Informatics, Vol. 131, No. 4, pp. 1-15, 2023.
5. S. S. Suresh, A. S. S. R. K. Prasad, and S. S. S. R. K. Prasad, "Optimized Machine Learning for the Early Detection of Polycystic Ovary Syndrome" **Sensors** (MDPI), Vol. 25, No. 4, 2025