ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 4| April 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

PREDICTIVE ANALYSIS OF PCOS USING CatBoost AND XGBoost: A MACHINE LEARNING APPROACH FOR EARLY DIAGNOSIS

Dr.N.Sathya[1], Sariga N[2], Meianbu B[3], Vishal S[4]
[1]*Associate Professor, Department of Artificial Intelligence and Machine Learning, Dr.N.G.P.Arts and Science College, Coimbatore, India.*
[2,3,4]*III B.Sc. AIML, Department of Artificial Intelligence and Machine Learning, Dr.N.G.P.Arts and Science College, Coimbatore, India.*

ABSTRACT
Polycystic Ovary Syndrome (PCOS) is a frequent endocrine illness in women of reproductive age, often leading to infertility, metabolic disorder, and cardiovascular disease. Early and appropriate diagnosis is critical for effective management, but routine diagnostic methods by clinical, biochemical, and ultrasonographic features are time-consuming and prone to variability. Machine learning (ML) presents a promising solution to enhance PCOS detection by identifying intricate patterns in patient data. This research utilizes XGBoost and CatBoost, two efficient gradient-boosting models, to create predictive models for PCOS diagnosis. The models are trained using a dataset of clinical, biochemical, and lifestyle features. Feature selection methods are utilized to determine the most important predictors, maximizing model performance. The models are tested using accuracy, precision, recall, and F1-score to ensure their reliability in clinical use.Experimental results demonstrate that XGBoost and CatBoost are better than classical models of classification with greater accuracy and stability of prediction in PCOS.

KEYWORDS: *PCOS, Machine Learning, Predictive Analysis, Early Diagnosis, Healthcare Analytics.*

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a common endocrine condition among reproductive-aged women that results in symptoms of infertility, hormonal disorders, obesity, and metabolic abnormalities. It also presents with long-term health risks, such as insulin resistance, type 2 diabetes, and cardiovascular disease.Although its extensive prevalence, PCOS is not easy to diagnose due to the variable signs and absence of an overarching diagnostic criterion.Traditional diagnostic methods such as clinical examination, biochemical tests, and ultrasound are time-consuming, expensive, and vulnerable to subjectivity.Machine learning (ML) has been shown to be a practical solution for the improvement of PCOS diagnosis from the exploration of complex patterns in clinical data. ML processes can deal with large amounts of data, identify chief predictors, and optimize diagnostic validity with the possibility of misdiagnosis minimized. This study explores the use of XGBoost and CatBoost, two efficient gradient-boosting models, in constructing predictive models for PCOS diagnosis. The predictive models rely on patient data, including clinical, biochemical, and lifestyle traits, to improve early detection and allow for individualized treatment strategies. By integrating ML techniques into PCOS diagnosis, this research aims to enhance prediction accuracy, lower diagnostic delay, and enhance health outcomes.

II. LITERATURE SURVEY

Over the past decade, application of machine learning (ML) methods in clinical practice has significantly enhanced predictive analysis of Polycystic Ovary Syndrome (PCOS). In a systematic review conducted in 2023 by the National Institutes of Health, the usefulness of AI and ML in the detection and diagnosis of PCOS accurately was found to be significant, as they can enable early detection and intervention treatment. In 2022, Phone, a specialist Convolutional Neural Network, was introduced by researchers. Phone was applied exclusively for detecting PCOS from ultrasound-derived ovarian images.Other advancements were brought to life in a 2024 paper describing Cyst Net, an artificial intelligence model with the use of multilevel threshold on ultrasound scans in detecting PCOS.

III METHODOLOGY

A. PROBLEM STATEMENT

The study aims to classify individuals into two categories:

- PCOS Positive
- PCOS Negative

The goal is to develop an efficient, accurate, and automated system using machine learning algorithms (CatBoost and XGBoost) to predict PCOS based on clinical, biochemical, and ultrasound features.

B. DATASET

The data used in Predictive Analysis of PCOS Using Machine Learning is derived from multiple sources. The Kaggle dataset (541 Excel records) has clinical, biochemical, and ultrasound attributes collected from research centers and hospitals. It provides patient data like BMI, hormone level, number of ovarian cysts, and menstrual irregularity. Moreover, a dataset of 2000+ train images and 2000+ test images with ultrasound images is used for deep learning-based PCOS detection. These datasets are predominantly obtained from medical research centers, Kaggle datasets, and hospital databases to design an

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 4| April 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

accurate and automated system to predict PCOS with the help of machine learning and deep learning methodologies.

C. DATA PREPROCESSING

It is a critical step to make the dataset clean, organized, and ready for machine learning models. Data cleaning is done first by managing missing values through mean imputation for numerical attributes and mode imputation for categorical attributes. Feature encoding is used to transform categorical variables such as menstrual irregularities and acne into numerical form through Label Encoding. Feature scaling is accomplished through Min-Max Normalization to convert all numerical attributes into a uniform range. Detection and management of outliers are facilitated with the use of the IQR method. The last steps include data balancing methodologies like SMOTE (Synthetic Minority Oversampling Technique) for handling class imbalance, producing an unbiased well-trained model.

IV. MODEL SELECTION – CATBOOST & XGBOOST

A. CatBoost (Categorical Boosting)

CatBoost is a Yandex-developed gradient boosting algorithm designed to perform categorical data in an efficient way. It's a decision tree-based and highly optimized for both speed during training and accuracy. CatBoost also distinguishes from other boosting algorithms by directly addressing categorical features and not requiring anyone hot or label encoding to prevent overfitting and gain performance.

Formula for CatBoost Model:

$$L=i=1\sum Nl(yi,F(xi))$$

where:

- L is the total loss
- yi is the actual value
- F(xi) is the predicted value
- l(yi,F(xi)) is the loss function (e.g., log loss for classification)

B. XGBoost (Extreme Gradient Boosting)

XGBoost is a boosted gradient algorithm optimized for increased speed and performance in dealing with large datasets. It is extensively applied in classification and regression problems because it can minimize overfitting, deal with missing values, and offer parallel processing for faster computation.

Formula for XGBoost Model:

$$L=i=1\sum Nl(yi\ ,y\hat{}i)+k=1\sum K\Omega(fk)$$

where:

- l(yi,y^i) is the loss function (e.g., squared error for regression, log loss for classification)
- Ω(fk)\Omega(f_k)Ω(fk) is the regularization term.
- fk represents the decision trees.

C. Model Evaluation

To assess model performance, we use:

- Accuracy = TP + TN / TP + TN + FP + FN
- Precision & Recall
  Precision = TP / TP + FP , Recall = TP / TP + FN
- F1-Score = 2 × Precision × Recall / Precision × Recall

V. EXPERIMENTAL RESULT

A. CatBoost

True Labels \ Predicted Labels	Predicted: No PCOS (0)	Predicted: PCOS (1)
Actual: No PCOS (0)	50 (True Negative - TN)	2 (False Positive - FP)
Actual: PCOS (1)	3(False Negative - FN)	64(True Positive - TP)

Table1: Confusion Matrix For Catboost

- True Positive (TP) = 63 → Correctly identified as PCOS.
- True Negative (TN) = 50 → Correctly identified as non-PCOS.
- False Positive (FP) = 2 → Misclassified as PCOS when they do not have PCOS.
- False Negative (FN) = 3 → Misclassified as non-PCOS when they actually have PCOS.

METRIC	RESULT
Accuracy	95.76%
Precision (PPV)	96.92%
Recall (Sensitivity)	95.45%
Specificity	96.15%
F1-Score	96.18%

Table 2: Performances Matrices For CatBoost

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 4| April 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

B. CATBOOST CONFUSION MATRIX

Figure:1 represents the confusion matrix of a CatBoost classifier, showing the model's performance in classifying two classes with 50 true negatives, 63 true positives, 2 false positives, and 3 false negatives.

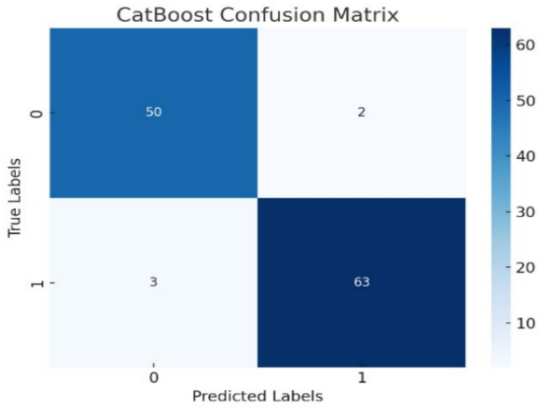

Figure:1

C. XGBoost

True Labels \ Predicted Labels	Predicted: No PCOS (0)	Predicted: PCOS (1)
Actual: No PCOS (0)	40 (True Negative - TN)	8 (False Positive - FP)
Actual: PCOS (1)	10 (False Negative - FN)	34 (True Positive - TP)

Table 3: Confusion Matrix for XGBoost

- TN (True Negative) = 40 → Correctly predicted non-PCOS cases.
- FP (False Positive) = 8 → Incorrectly predicted PCOS when it was non-PCOS.
- FN (False Negative) = 10 → Incorrectly predicted non-PCOS when it was PCOS.
- TP (True Positive) = 34 → Correctly predicted PCOS cases.

D. PERFORMANCES MATRICES

METRIC	RESULT
Accuracy	82.76 %
Precision (PPV)	80.95 %
Recall (Sensitivity)	77.27%
Specificity	83.33%
F1-Score	79.07%

Table 4: Performances Matrices for XGBoost

E. XGBOOST CONFUSION MATRIX

Figure2 represents the confusion matrix of an XGBRF classifier, showing the model's classification performance with 40 true negatives, 34 true positives, 8 false positives, and 10 false negatives.

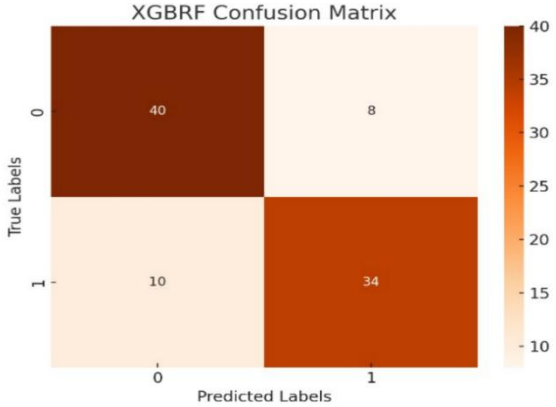

Figure: 2

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 4| April 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

VI. COMPARISON OF CATBOOST AND XGBOOST

CatBoost and XGBoost are both strong gradient boosting algorithms but differ in their processing of categorical data, computational capabilities, and precision. According to the provided bar graph, CatBoost is far more precise (approximately 93.5%) than XGBRF (approximately 85.5%), which indicates its superior performance.CatBoost, created by Yandex, is designed to efficiently process categorical features without explicit encoding, cutting down on preprocessing. It implements ordered boosting, minimizing overfitting and bias in small data, making it especially suitable for situations involving high-cardinality categorical data. Its implementation also facilitates efficient training and inference while maintaining stability on different datasets.XGBoost, however, is a scalable and fast gradient boosting library that is popular. XGBRF (XGBoost with Random Forest method), however, is problematic with categorical data, where one-hot encoding has to be applied, which is more complex and computationally expensive.

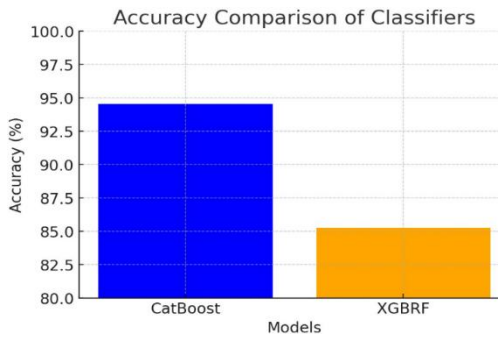

Figure:3 represents an accuracy comparison of classifiers, showing that CatBoost outperforms XGBRF in classification accuracy.

VII. CONCLUSION

The findings of this study emphasize the potential of AI-driven healthcare solutions in diagnosing complex medical conditions like PCOS. The Predictive Analysis of PCOS Using Machine Learning study illustrates the capability of sophisticated algorithms in early diagnosis and detection of Polycystic Ovary Syndrome (PCOS). By leveraging machine learning, medical professionals can enhance diagnostic efficiency, reduce the dependency on invasive tests, and improve patient outcomes. With patient data, machine learning algorithms can identify patterns related to the condition so that intervention can be made early. The comparison of CatBoost and XGBoost shows that the two models are both excellent, but that CatBoost is slightly better at classification. The research highlights the importance of feature selection, data preprocessing, and correct model evaluation in giving accurate output. The research paves the way for AI-based diagnostic solutions and reduces reliance on invasive testing and enhances clinical decision-making. Further research can focus on integrating deep learning techniques and expanding datasets to further refine predictive accuracy.

REFERENCES

1. *Bharati et al. (2020) applied various machine learning algorithms, including Decision Trees and Support Vector Machines, achieving high accuracy, highlighting the potential of these models in diagnosing PCOS.*
2. *Denny et al. (2019) developed i-hope, a detection and prediction system for PCOS using machine learning techniques, demonstrating the effectiveness of integrated approaches in early diagnosis.*
3. *Anda and Iyamah (2022) conducted a comparative analysis of artificial intelligence in the diagnosis of PCOS, emphasizing the role of AI in enhancing diagnostic accuracy.*
4. *Bhardwaj and Tiwari (2022) explored the application of machine learning algorithms in the healthcare sector, specifically for PCOS diagnosis, showcasing the potential of these technologies in medical diagnostics.*
5. *Adla et al. (2021) implemented automated detection of PCOS using machine learning techniques, highlighting the efficiency of automated systems in medical diagnostics.*
6. *Thakre et al. (2020) developed PCOcare, a system for PCOS detection and prediction using machine learning algorithms, demonstrating the practical application of these models in healthcare.*
7. *Chauhan et al. (2021) conducted a comparative analysis of machine learning algorithms for PCOS prediction, providing insights into the most effective models for this application.*
8. *Rathod et al. (2022) utilized the CatBoost algorithm for predictive analysis of PCOS, achieving notable accuracy, and highlighting the algorithm's effectiveness in handling complex datasets.*
9. *Aggarwal et al. (2023) proposed an improved technique for PCOS risk prediction using feature selection and machine learning, emphasizing the importance of relevant feature selection in enhancing model performance.*
10. *Khanna et al. (2023) developed a distinctive explainable machine learning framework for PCOS detection, focusing on the interpretability of AI models in medical diagnostics.*