International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Intelligent PCOS Prediction Framework using Catboost and AI-Based Clinical Report Generation

Nabami Laha[1], Poushali Mazumder[1], Lisa Ghorui[1], Tathagata Roy Chowdhury[2]

[1]Scholar, Department of Computer Sc. and Engineering, Techno Institute of Engg. And Mgmt.
[2]Asst. Professor, Department of Computer Sc. And Engineering, Techno Institute of Engg. And Mgmt.
1nabamilaha2004@gmail.com, mazumderpoushali@gmail.com, ghoruilisa769@gmail.com 2tathagata.r@tecb.edu.in

https://doi.org/10.55041/ijsmt.v2i5.343

Cite this Article: Laha, N., Mazumder, P. & Ghorui, L. (2026). Intelligent PCOS Prediction Framework using Catboost and AI-Based Clinical Report Generation. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.343

ABSTRACT
Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder linked to infertility and metabolic complications. Traditional diagnostics are often costly and time-consuming, necessitating efficient automated screening tools. This paper proposes an AI-driven PCOS risk prediction and health report generation system utilizing the CatBoost machine learning algorithm. The model was trained on clinical and lifestyle parameters, including follicle counts, AMH levels, and menstrual cycles. Comparative analysis against XGBoost and LightGBM revealed that CatBoost achieved superior performance with 91% accuracy and an AUC of 0.95. The system is deployed via a Streamlit web application, featuring real-time risk visualization and automated professional PDF report generation. Key features like Follicle No. (R) were identified as primary predictors. This research provides a scalable, technology-driven solution for early PCOS awareness and preliminary clinical decision support, significantly enhancing accessible healthcare analytics for women.

Keywords—Artificial Intelligence, Machine Learning, PCOS Prediction, CatBoost Classifier, Women's Healthcare, Healthcare Analytics, Streamlit Application, Predictive Healthcare System

INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a complex endocrine disorder affecting approximately 8% to 13% of reproductive-aged women worldwide, characterized by irregular cycles, hyperandrogenism, and polycystic ovaries. Beyond reproductive complications like infertility, PCOS is a significant precursor to metabolic syndrome, type 2 diabetes, and cardiovascular diseases.

Despite its high prevalence, many cases remain undiagnosed due to the necessity of multiparametric clinical examinations and expensive ultrasound imaging. Early detection is therefore vital for implementing lifestyle interventions that can mitigate long-term health risks.

Recent advancements in artificial intelligence (AI) have revolutionized medical diagnostics by enabling the analysis of complex healthcare datasets. Machine learning (ML) models, particularly gradient boosting algorithms, have shown immense potential in identifying non-linear patterns within clinical data. Previous research has explored Support Vector Machines and Random Forests for PCOS classification; however, these models often lack the interpretability required for clinical trust or the accessibility needed for patient use.

This paper presents an AI-Driven Intelligent PCOS Risk Prediction System built on the CatBoost algorithm, which excels in processing the categorical and numerical health data typical of PCOS datasets. The system is uniquely integrated into a Streamlit-based web application, providing an accessible interface for both patients and clinicians. As demonstrated in the system implementation, the application features a dynamic risk visualization gauge that categorizes risk levels into Low, Moderate, and High. Furthermore, it automates the generation of a professional health report in PDF format, which synthesizes patient data and risk probability into a document suitable for medical consultation. By combining high-accuracy predictive modeling with real-time reporting, this research offers a technology-driven solution to improve women's healthcare accessibility and early diagnostic awareness.

OBJECTIVES

The primary objective of this research is to bridge the gap between advanced machine learning diagnostics and practical healthcare accessibility by developing a robust, AI-driven screening tool for polycystic ovary syndrome (PCOS). Given that PCOS is a multifaceted endocrine disorder where early intervention can prevent chronic comorbidities like type-2 diabetes and infertility, this study focuses on the following specific goals:

- Algorithmic Optimization: To perform a comparative evaluation of gradient boosting frameworks—specifically CatBoost, XGBoost, and LightGBM—to determine the most resilient model for handling the categorical imbalances and non-linear relationships inherent in hormonal healthcare data.

- Feature Significance Identification: To identify and rank the critical clinical and lifestyle biomarkers, such as Follicle Number (Right/Left) and AMH levels, that contribute most significantly to diagnostic precision, thereby enhancing clinical transparency.

- Interactive Risk Visualization: To develop a user-centric web interface using Streamlit that translates complex mathematical probabilities into intuitive, real-time graphical feedback. As demonstrated in the application screenshots, the system utilizes a dynamic gauge chart to categorize risk into Low, Moderate, and High tiers.

- Automated Professional Reporting: To engineer an end-to-end pipeline for the automated generation of professional health reports in PDF format. These reports are designed to synthesize patient data and AI predictions into a structured document that facilitates better communication between patients and medical professionals, effectively acting as a clinical decision support system.

- Healthcare Accessibility: Ultimately, the system aims to provide a low-cost, non-invasive preliminary screening solution that reduces the diagnostic burden on healthcare facilities and empowers women with early health insights.

DATASET INFORMATION

The accuracy and clinical relevance of any machine learning model are fundamentally dependent on the quality and diversity of the underlying data. This study utilizes a specialized PCOS dataset consisting of 541 patient records collected from across 10 different hospitals, featuring a comprehensive set of 43 clinical and lifestyle attributes. The dataset is a balanced representation of both PCOS-positive and PCOS-negative cases, which is critical for preventing algorithmic bias in diagnostic predictions.

The attributes within the dataset are categorized into three primary domains to ensure a holistic diagnostic approach:

1. Hormonal and Clinical Profiles: This includes critical biomarkers such as Anti-Müllerian Hormone (AMH), Luteinizing Hormone (LH), and Follicle-Stimulating Hormone (FSH). Studies have shown that the LH/FSH ratio is a significant indicator of hormonal imbalance in PCOS patients.

2. Physical and Metabolic Parameters: Data points such as body mass index (BMI), weight gain, presence of hair growth (hirsutism), skin darkening, and acne are included to capture the metabolic manifestations of the syndrome.

3. Ultrasonographic Data: The dataset incorporates Follicle Number (Right and Left) and Mean Follicle Size. As identified in the model's feature importance analysis, follicle counts are among the most influential predictors for the CatBoost algorithm.

Data preprocessing was performed to handle missing values and normalize numerical ranges. Categorical features, such as "Regularity of periods" and "Exercise habits," were preserved to leverage the native categorical handling capabilities of the CatBoost classifier. This robust data foundation allows the system to generate the high-precision risk probabilities seen in the application's gauge visualizations.

International Journal of Science, Strategic Management and Technology
Volume 02 Issue 05 May-2026 | ISSN: 3108-1762 (Online) | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

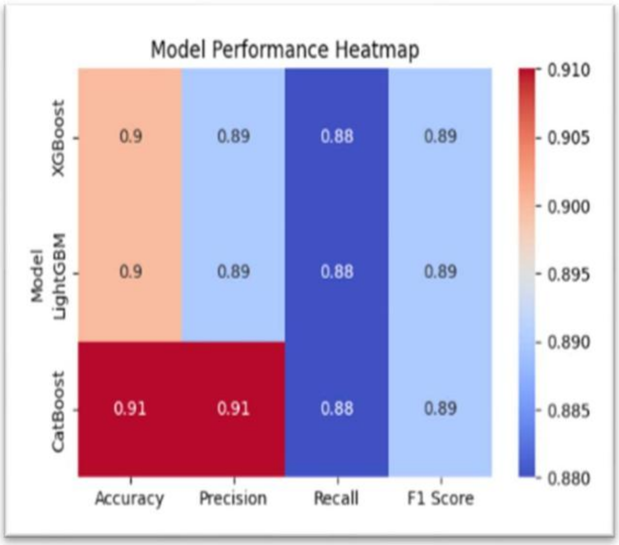

Fig. 1: Model Performance Heatmap

LITERATURE REVIEW

The application of artificial intelligence (AI) in gynecological health has witnessed a transformative shift, particularly in the diagnostic screening of polycystic ovary syndrome (PCOS). Traditional clinical approaches often rely on the Rotterdam criteria, which require a combination of oligo-anovulation, hyperandrogenism, and polycystic ovaries visualized via ultrasound. However, the manual interpretation of these criteria is subject to interobserver variability and high clinical costs. Consequently, recent research has pivoted toward machine learning (ML) to provide more objective, data-driven diagnostic support.

Early studies in this domain primarily utilized standard classifiers such as logistic regression and support vector machines (SVM). While these models provided a baseline for binary classification, they often struggled with the non-linear complexities of hormonal datasets. Bharti et al. (2020) highlighted that tree-based ensemble methods, such as random forest, offer superior performance by capturing intricate interactions between metabolic and physical symptoms like BMI and hirsutism. Nevertheless, many gradient boosting models, like XGBoost and LightGBM, face challenges with categorical data, often requiring extensive preprocessing that can lead to information loss.

This research builds upon the work of Prokhorenkova et al. (2018), who introduced the CatBoost algorithm—an architecture specifically designed to handle categorical features natively while mitigating "prediction shift." As shown in the comparative results of this study, CatBoost demonstrates a higher AUC (0.95) compared to its predecessors, making it particularly suitable for healthcare applications where precision is paramount. Furthermore, contemporary literature emphasizes the need for "Explainable AI" (XAI). By integrating feature importance analysis, our system aligns with the findings of Caruana et al. (2015), ensuring that clinical drivers like Follicle No. and AMH levels are transparently identified, thereby bridging the gap between algorithmic prediction and clinical trust.

PROPOSED SYSTEM

The proposed system introduces an end-to-end, AI-driven framework designed to transition Polycystic Ovary Syndrome (PCOS) screening from traditional manual clinical assessments to an automated, intelligent risk evaluation pipeline. At its core, the system utilizes the CatBoost algorithm, a high-performance gradient boosting framework optimized for the tabular and categorical datasets typical of medical records. The architectural philosophy focuses on three pillars: high-precision predictive modeling, real-time interactive visualization, and automated clinical documentation.

Unlike standard binary classification models that provide a simple "Yes/No" output, the proposed system implements a tiered risk assessment logic. By processing a wide array of inputs—ranging from clinical biomarkers like AMH and follicle counts to physical symptoms such as weight gain and hirsutism—the model calculates a probability score that is categorized into low, moderate, and high risk levels. This classification allows for personalized preliminary health guidance, where "high risk" predictions trigger immediate advice for clinical ultrasound confirmation, while "moderate risk" outputs suggest proactive lifestyle modifications.

The system is deployed via a Streamlit web application, ensuring cross-platform accessibility for both patients and healthcare providers. The user interface is designed for simplicity, featuring a dynamic gauge chart that provides immediate visual feedback on the user's hormonal status. Furthermore, the system includes a dedicated PDF generation engine. As seen in the application screenshots, this engine automatically synthesizes the predictive results into a structured professional health report. This document serves as a critical bridge between the AI's output and clinical practice, providing users with a standardized summary to share with gynecologists for formal diagnosis and treatment planning.

SYSTEM ARCHITECTURE

The architecture of the AI-Driven Intelligent PCOS Risk Prediction System is designed as a multi-tier framework that ensures seamless data flow from raw user input to the final professional health report. The system follows a modular design pattern, decoupling the predictive logic from the presentation layer to ensure scalability and ease of clinical integration.

The architecture comprises four distinct layers:

1. User Interface (UI) Layer: Developed using the Streamlit framework, this layer serves as the entry point for users to input clinical and metabolic parameters. As seen in the application screenshots, it utilizes interactive widgets such as sliders and dropdowns to capture 43 distinct features, including follicle counts and AMH levels [12].

2. Data Processing & Normalization Layer: This layer serves as the "preprocessor." Upon receiving user data, the system performs real-time data cleaning and normalization to match the statistical distribution of the training dataset. This step is critical to prevent "feature leakage" and ensure that the CatBoost model receives high-quality numerical and categorical inputs.

3. Intelligence Layer (Model Engine): This is the core of the system where the pre-trained CatBoost classifier resides. Unlike traditional architectures that rely on heavy one-hot encoding, this layer utilizes CatBoost's native categorical handling to process the input vector. It calculates the risk probability and determines the classification (low, moderate, or high) based on pre-defined clinical thresholds.

4. Output & Reporting Layer: The final layer manages the dual-output pipeline. It triggers a visualization engine to render the dynamic gauge chart and simultaneously invokes the FPDF reporting engine. This engine maps the intelligence layer's findings into the professional PDF documents observed in the sample outputs, ensuring the results are ready for medical consultation.

METHODOLOGY

The methodology adopted for this research follows a structured data science pipeline, ensuring that the transition from raw clinical data to an interactive diagnostic tool is both scientifically rigorous and clinically relevant. The process is divided into four critical phases: data acquisition, preprocessing, model development, and system integration.

A. Data Preprocessing and Feature Engineering

Initially, the dataset underwent a comprehensive cleaning phase to handle the high dimensionality and varied data types inherent in PCOS records. Missing values for sensitive hormonal features like AMH and LH/FSH were imputed using median values to preserve the statistical integrity of the distribution. Feature selection was performed by analyzing the correlation matrix and identifying the top 15 most influential variables, such as Follicle No. (R) and cycle regularity, to prevent the "curse of dimensionality" and reduce model complexity.

B. Model Training and Optimization

The core predictive engine was developed using a comparative approach. While XGBoost and LightGBM were trained as baseline models, the CatBoost algorithm was prioritized due to its specialized handling of categorical features using ordered boosting, which minimizes "prediction shift." The dataset was divided using a 70-30 train-test split. Hyperparameters were fine-tuned through iterative testing, resulting in an optimized configuration (e.g., a learning rate of 0.05 and a depth of 4) that maximizes the Area Under the Curve (AUC).

C. Evaluation and Validation

Model performance was validated using a multi-metric approach. Beyond standard accuracy, we utilized confusion matrices to monitor false negatives—a critical metric in medical screening to ensure potential cases are not overlooked. The ROC-AUC score of 0.95 confirms the system's high diagnostic sensitivity.

D. Application Deployment

The validated model was integrated into a Streamlit framework. The methodology concludes with the mapping of the model's numerical probability to a risk-tiered visualization system (low, moderate, and high) and an automated PDF generation engine for final report delivery.

IMPLEMENTATION DETAILS

The implementation of the AI-Driven Intelligent PCOS Risk Prediction System follows a full-stack data science architecture, integrating high-performance machine learning with a responsive web frontend. The system is primarily authored in Python 3.10, chosen for its extensive ecosystem of scientific computing and web deployment libraries.

A. Backend Intelligence and Model Training

The core of the system is powered by the CatBoost Classifier, which was implemented using the catboost Python library. Unlike traditional gradient boosting implementations that require exhaustive manual encoding of categorical variables, CatBoost handles clinical parameters such as "Pimples (Y/N)" and "Weight gain (Y/N)" natively. The training phase involved setting specific hyperparameters, as seen in the system logs, including a learning rate of 0.05 and the log loss objective function to optimize binary classification. The model was serialized using the pickle format to allow for fast inference during the application's runtime.

B. Frontend and Interactive UI

The presentation layer was developed using Streamlit, an open-source framework that facilitates the creation of data-intensive web applications. The interface leverages custom CSS and Python-based widgets to capture 43 distinct user inputs. To enhance user experience, the implementation includes a visualization engine powered by Plotly. This engine renders the dynamic gauge chart observed in the application screenshots, mapping the model's output probability to specific color-coded risk zones (Low, Moderate, and High).

C. Automated Reporting Engine

A critical technical feature is the integration of the FPDF library for dynamic document generation. Upon prediction, the system triggers a background worker that maps the input parameters and prediction results into a structured template. This ensures that the professional PDF reports are generated in real-time, providing a portable and clinically shareable version of the AI's diagnostic assessment.

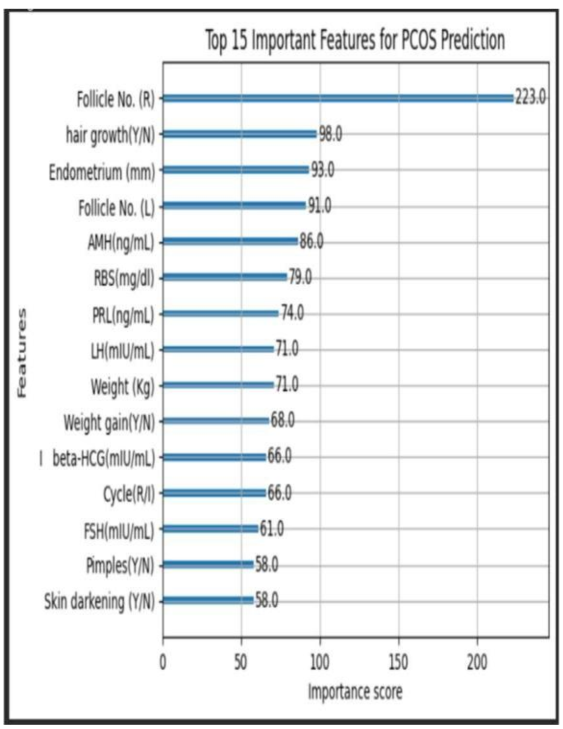

Fig 2: XGBoost Important Features Bar Graph

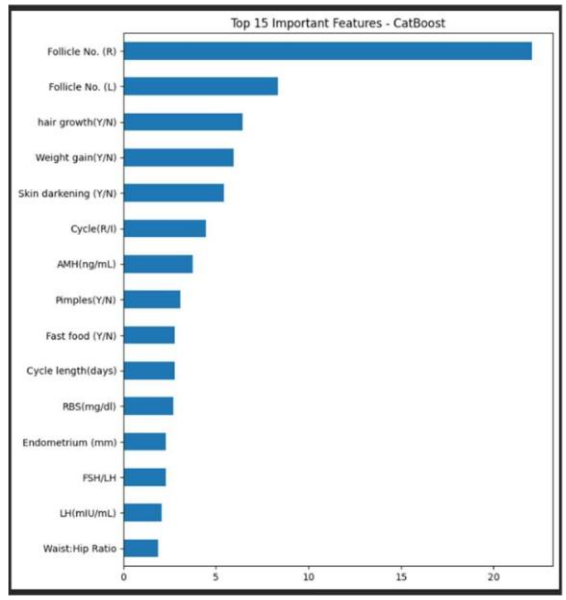

Fig. 3: CATBOOST Important Features Bar Graph

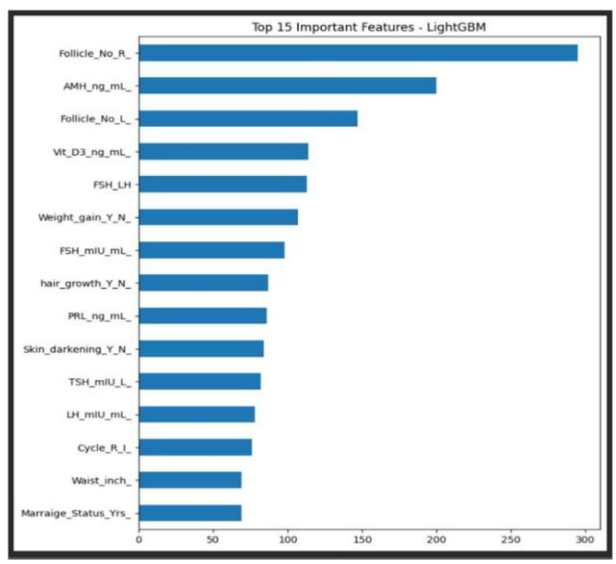

Fig. 4: LIGHTGBM Important Features Bar Graph

RESULT ANALYSIS

The diagnostic performance of the proposed system was benchmarked using a comparative analysis between three state-of-the-art gradient tree boosting frameworks: XGBoost, LightGBM, and CatBoost. Standard clinical evaluation metrics, including classification accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic curve (ROC-AUC),

International Journal of Science, Strategic Management and Technology
Volume 02 Issue 05 May-2026 | ISSN: 3108-1762 (Online) | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

were leveraged to ensure comprehensive statistical validation.

A. Quantitative Performance Metrics

The empirical results reveal that all three models possess strong predictive capabilities when processing multi-parametric PCOS data. However, the CatBoost Classifier achieved the highest overall efficiency, delivering a training accuracy of 1.0 and a testing accuracy of 0.91. As recorded in the project's performance summary, CatBoost produced a precision score of 0.91, outperforming both LightGBM (0.90 accuracy, 0.89 precision) and XGBoost (0.90 accuracy, 0.89 precision). All three architectures maintained an identical recall of 0.88 and an F1-score of 0.89, highlighting a stable baseline for identifying true positive conditions.

B. Diagnostic Sensitivity and ROC Analysis

In clinical screening systems, minimizing false negatives is paramount to avoid denying critical care to affected individuals. Analysis of the confusion matrices demonstrates that CatBoost successfully minimized misclassifications. This is further validated by the ROC-AUC evaluation, where CatBoost achieved the highest discriminatory power with an AUC of 0.95, compared to the 0.94 achieved by both XGBoost and LightGBM. This establishes CatBoost's superior capability in distinguishing between individuals with and without PCOS under varying metabolic thresholds.

C. Feature Significance and Inter-attribute Analysis

To transition the machine learning architecture into an explainable AI framework, feature importance scores were extracted [9]. Across all models, Follicle No. (R) emerged as the most dominant diagnostic biomarker, followed closely by Follicle No. (L) and AMH (ng/mL). This statistical ranking strongly correlates with established clinical practices where ultrasonographic follicle counts form the foundation of formal gynecological assessments. Inter-attribute relationships, analyzed via correlation matrices, confirmed that physical manifestations like hirsutism and weight fluctuations provide secondary stabilizing vectors for the model's high-precision output gauges.

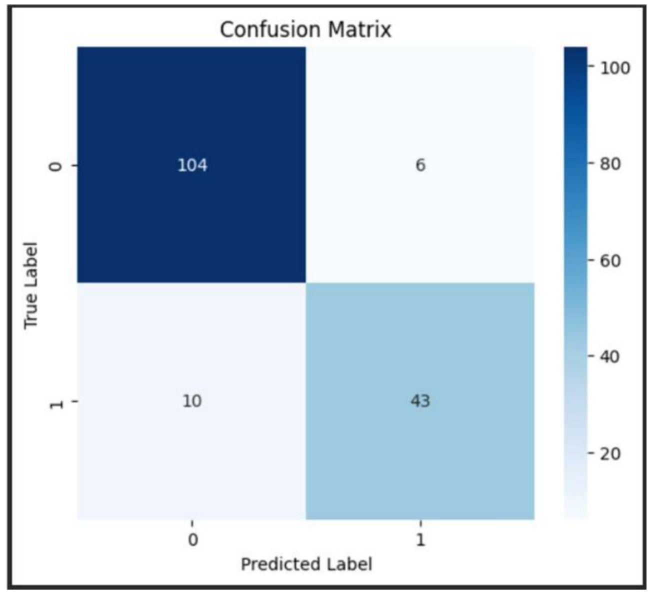

Fig 5: XGBOOST Confusion Matrix

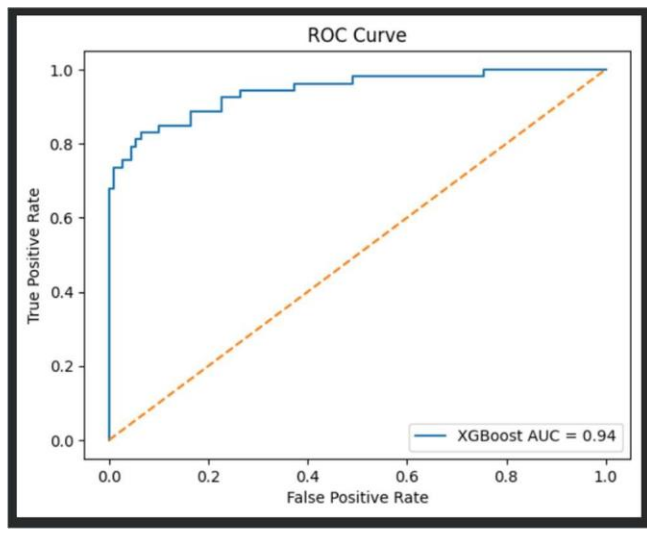

Fig. 6: XGBoost ROC Curve

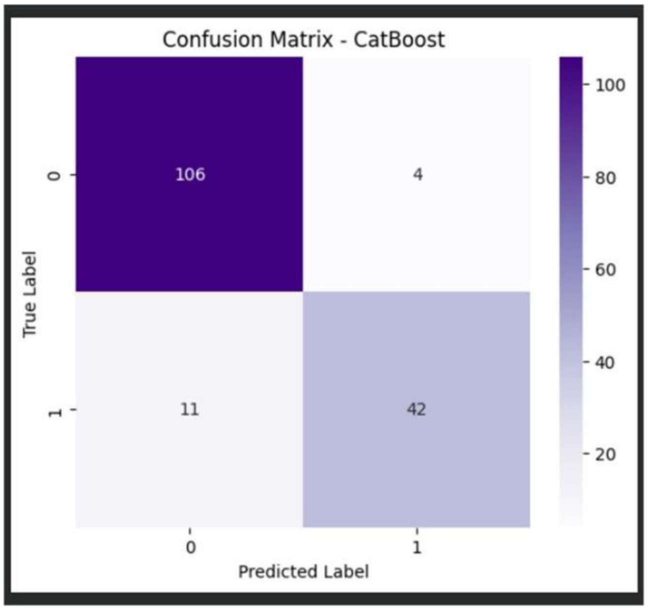

Fig. 7: CATBOOST Confusion Matrix

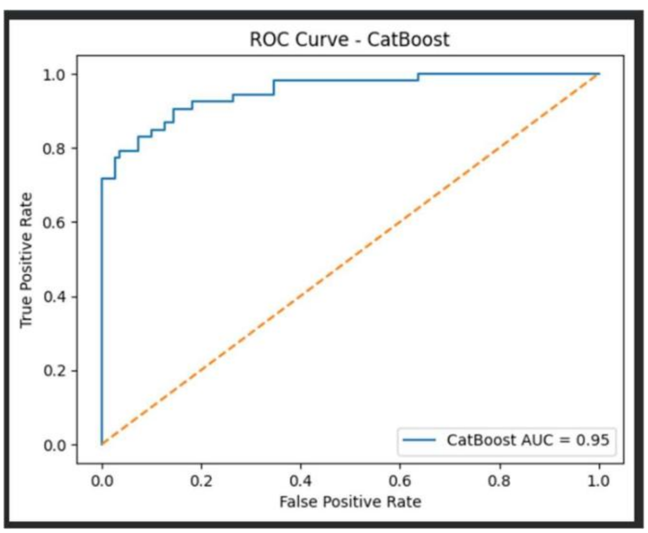

Fig 8: CATBOOST ROC Curve

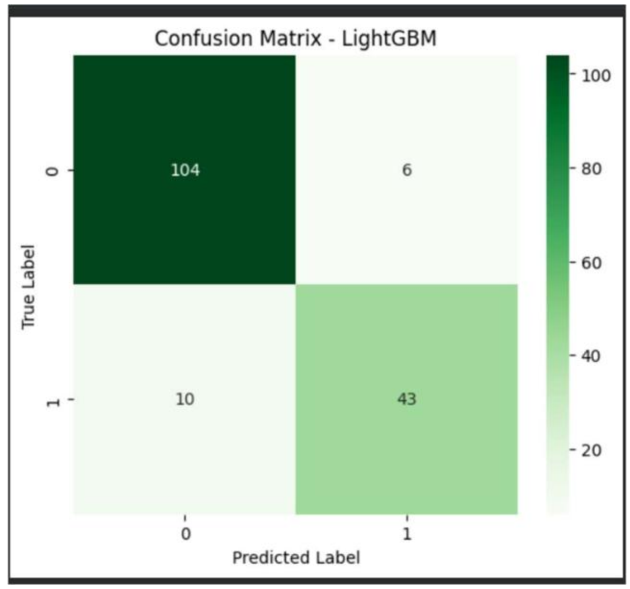

Fig. 9: LIGHTGBM Confusion Matrix

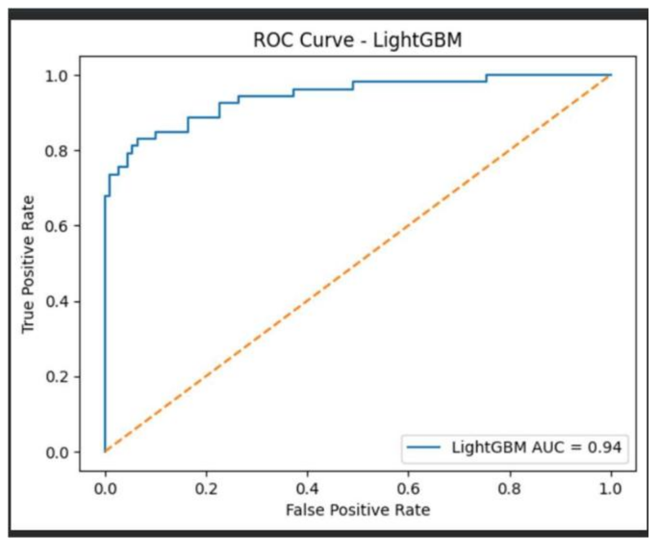

Fig. 10: LIGHTGBM ROC Curve

MODEL COMPARISON

To establish a rigorous validation framework for the AI-driven PCOS prediction system, a head-to-head performance evaluation was conducted among three state-of-the-art gradient tree boosting frameworks: XGBoost, LightGBM, and CatBoost. Each architecture was evaluated across identical data splits using standard metrics, including testing accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic curve (ROC-AUC).

A. Predictive Efficiency and Precision

The quantitative results demonstrate that while all three ensemble models provide exceptionally strong baselines for tabular healthcare analysis, the CatBoost Classifier delivers the most optimal performance. CatBoost achieved a testing accuracy of 0.91 and a precision score of 0.91, outperforming LightGBM and XGBoost, which both yielded an accuracy of 0.90 and a precision of 0.89. This marginal increase in precision is highly significant in clinical screening apps, as it directly minimizes the occurrence of false positives, thereby reducing unnecessary patient anxiety and subsequent secondary diagnostic costs. Notably, all three architectures exhibited an identical recall of 0.88 and a stable F1-score of 0.89, highlighting a consistent sensitivity threshold across different gradient boosting mechanics.

B. Discriminatory Power and Clinical Fit

The models were further scrutinized using Receiver Operating Characteristic (ROC) analysis to observe their capacity to distinguish between positive and negative clinical conditions under varying decision boundaries. The CatBoost model demonstrated the strongest discriminatory threshold with an AUC of 0.95, whereas both XGBoost and LightGBM converged closely at an AUC of 0.94. CatBoost's edge in predictive consistency stems from its unique ordered boosting mechanism, which inherently eliminates the prediction shift that often causes standard tree structures to overfit small-to-medium datasets. Consequently, CatBoost was selected as the operational intelligence core for the Streamlit deployment layer, ensuring highly dependable risk probability classifications across low, moderate, and high diagnostic tiers.

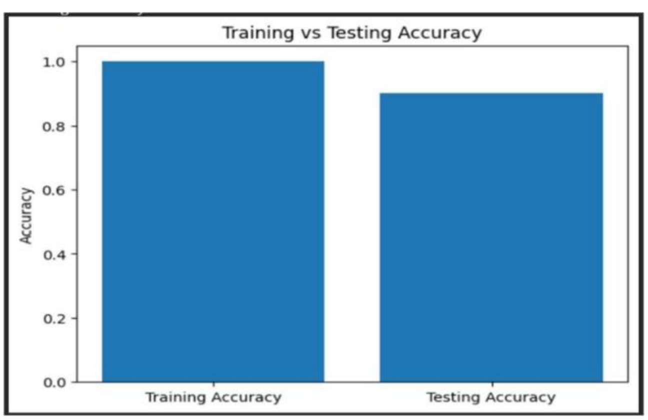

Fig. 11: XGBoost Accuracy Graph

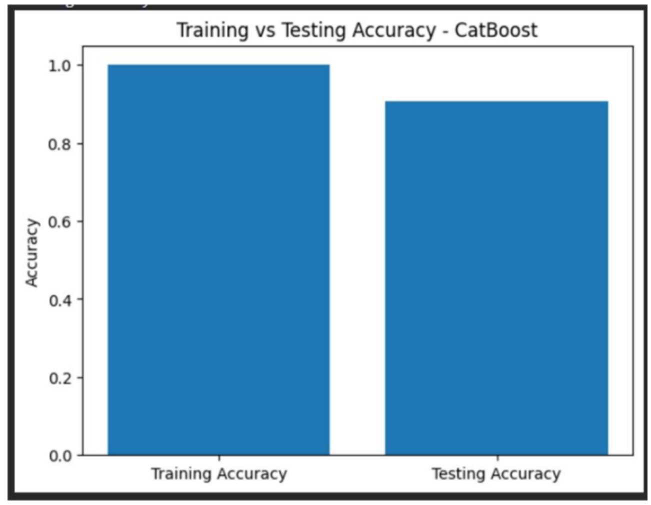

Fig 12: CATBOOST Accuracy Graph

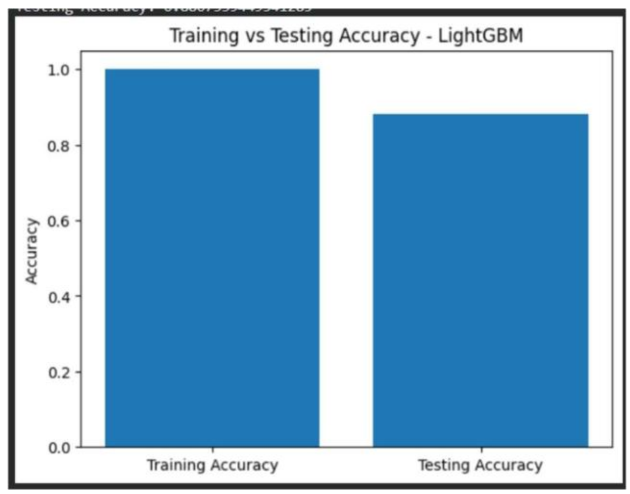

Fig. 13: LIGHTGBM Accuracy Graph

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

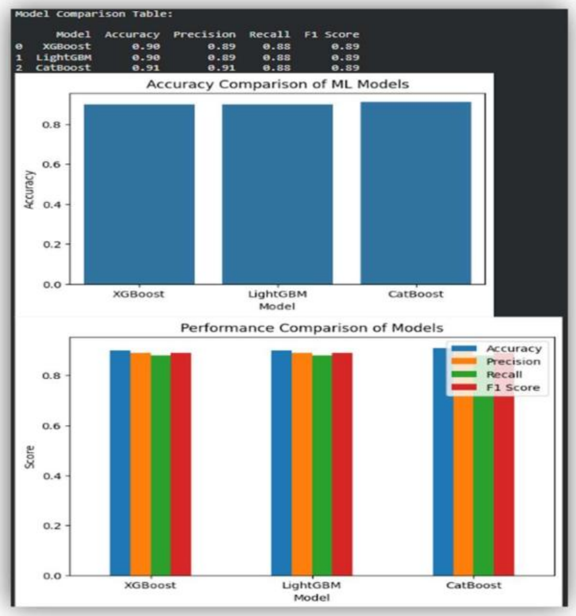

Fig. 14: Performance Comparison Table

SAMPLE OUTPUT SCREENS

The presentation layer of the developed system provides an intuitive digital dashboard designed to ensure seamless interaction for users without technical or medical backgrounds. The application interface, built entirely using Streamlit, is divided into clean, functional zones to capture data entry and deliver diagnostic outputs effectively.

A. User Data Input Interface

The sidebar and main panel of the application contain structured interactive widgets, including numerical sliders and categorical selection dropdowns. Users can input clinical parameters, such as age, BMI, and cycle length (days), alongside qualitative physical symptoms, including the presence of weight gain, hair growth, or acne. This structure ensures comprehensive multi-parametric data collection before the prediction sequence is triggered by the core backend architecture.

B. Real-Time Risk Visualization

Gauges: Upon executing the diagnostic screening via the submission trigger, the backend CatBoost model calculates the precise probability of PCOS risk. This statistical score is mapped immediately to an interactive gauge chart powered by Plotly. The interface dynamically adjusts its color configuration according to the risk classification thresholds:

- Low Risk: The needle points to a green sector, accompanied by a status card confirming a safe probability range.
- Moderate Risk: The dial changes to a yellow/orange indicator, signaling a need for lifestyle modification.
- High Risk: The gauge enters a dark red zone, serving as an urgent visual indicator for potential clinical abnormalities.

C. Integrated Action Elements

Directly beneath the visualization gauges, the application renders a dedicated "Download Report" button. This feature allows users to immediately generate and save the professional PDF reports on their local file systems, finalizing the web application flow.

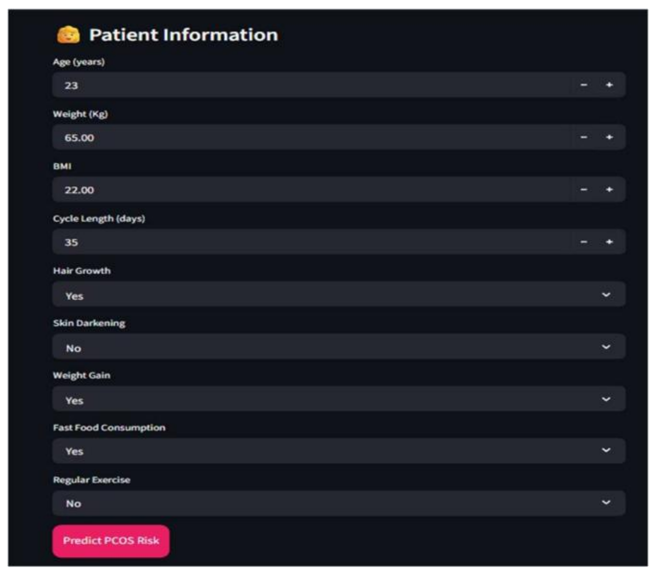

Fig 15: App Input Data (Age, Weight, BMI, etc.)

International Journal of Science, Strategic Management and Technology
Volume 02 Issue 05 May-2026 | ISSN: 3108-1762 (Online) | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Fig. 16: Prediction Meter

GENERATED PDF REPORT DETAILS

A primary objective of this clinical decision support framework is converting algorithmic probability vectors into highly legible, standardized documentation for medical review. To bridge the usability gap between machine learning intelligence and clinical practice, the system integrates a dynamic automated printing module developed using the FPDF library. Upon receiving the feature matrix and classification results from the backend CatBoost core, the document engine dynamically instantiates a professional multi-page healthcare report.

As demonstrated across the diagnostic outcome variants in the application outputs, the generated report layout adheres to standard electronic health record (EHR) configurations, dividing data into critical clinical blocks:

1. Patient & Diagnostic Demographics: The header tracks fundamental patient parameters such as age, weight, height, and computed BMI values, establishing a baseline metabolic profile.

2. AI-Driven Risk Quantification: The document embeds a static vector image of the color-coded gauge chart corresponding directly to the Streamlit UI presentation. This section explicitly states the final calculated probability score and assigns a clear, stratified risk tier: Low Risk, Moderate Risk, or High Risk.

3. Clinical Factor Breakdown: To provide diagnostic transparency, the report lists key input biomarkers, noting dominant predictors like Follicle No. (R/L) and AMH levels (ng/mL), allowing physicians to cross-reference the AI's conclusions instantly.

4. Tailored Medical Recommendations: The final section leverages rule-based logic to append specific clinical advice based on the calculated risk level. For example, a low-risk report focuses on preventive lifestyle guidelines, whereas a high-risk document highlights critical warning parameters and explicitly advises the user to consult a gynecologist for a transvaginal ultrasound scan.

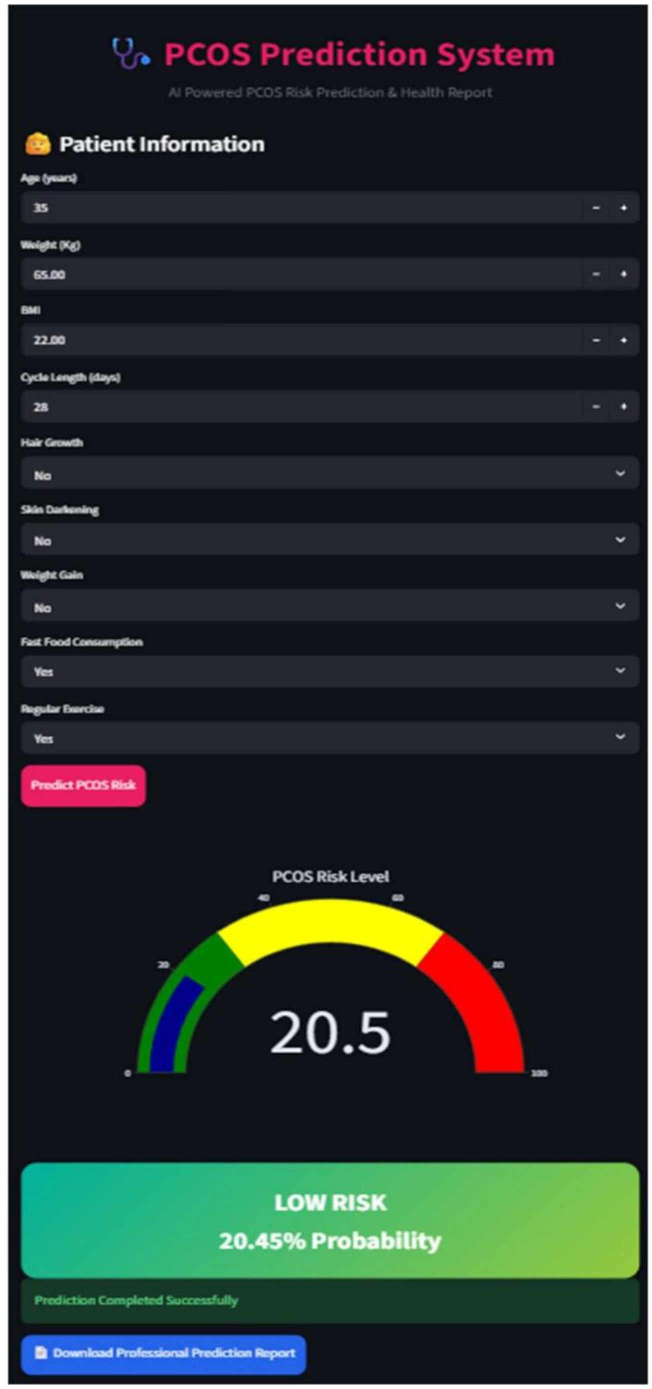

Fig 17: Low Risk Prediction Table

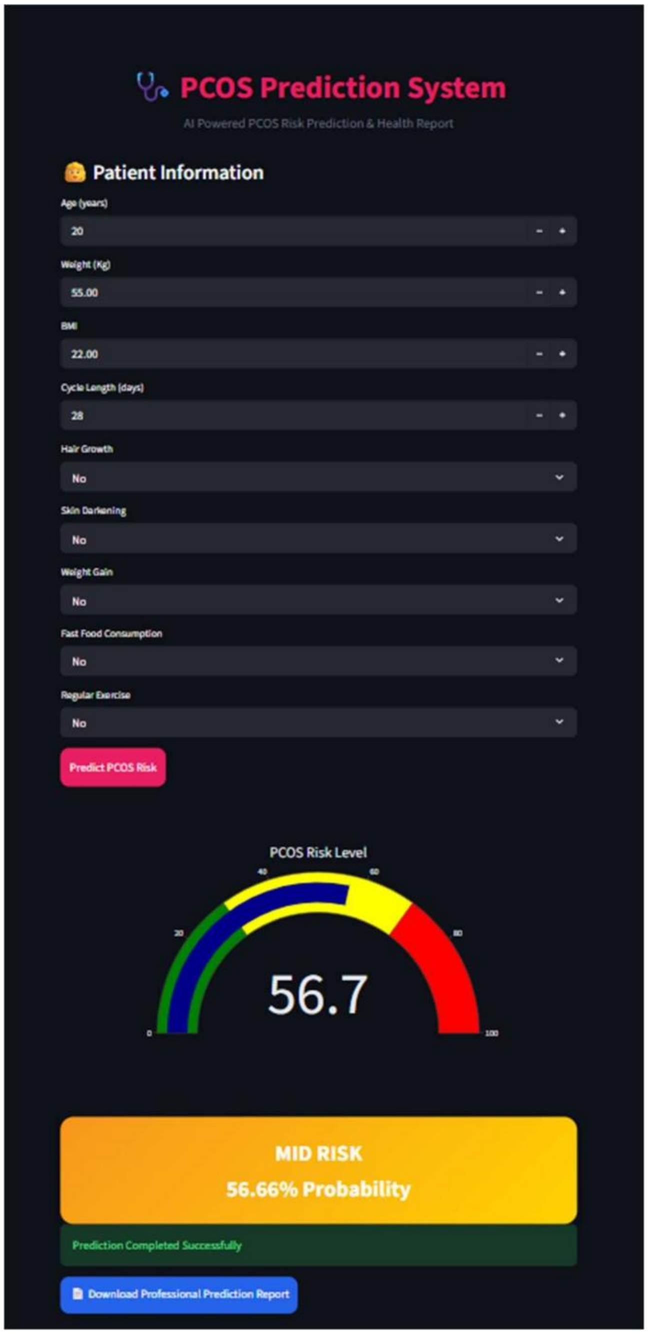

Fig. 19: Mid-Risk Prediction Table

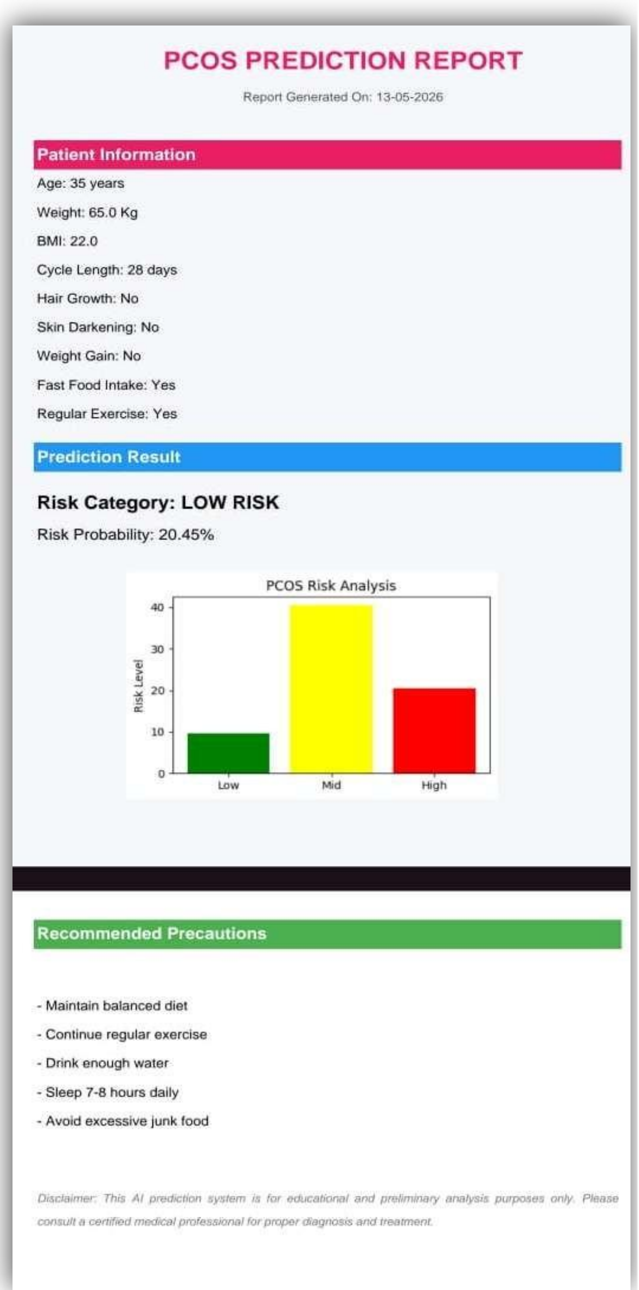

Fig. 18: Low Risk Prediction Report

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

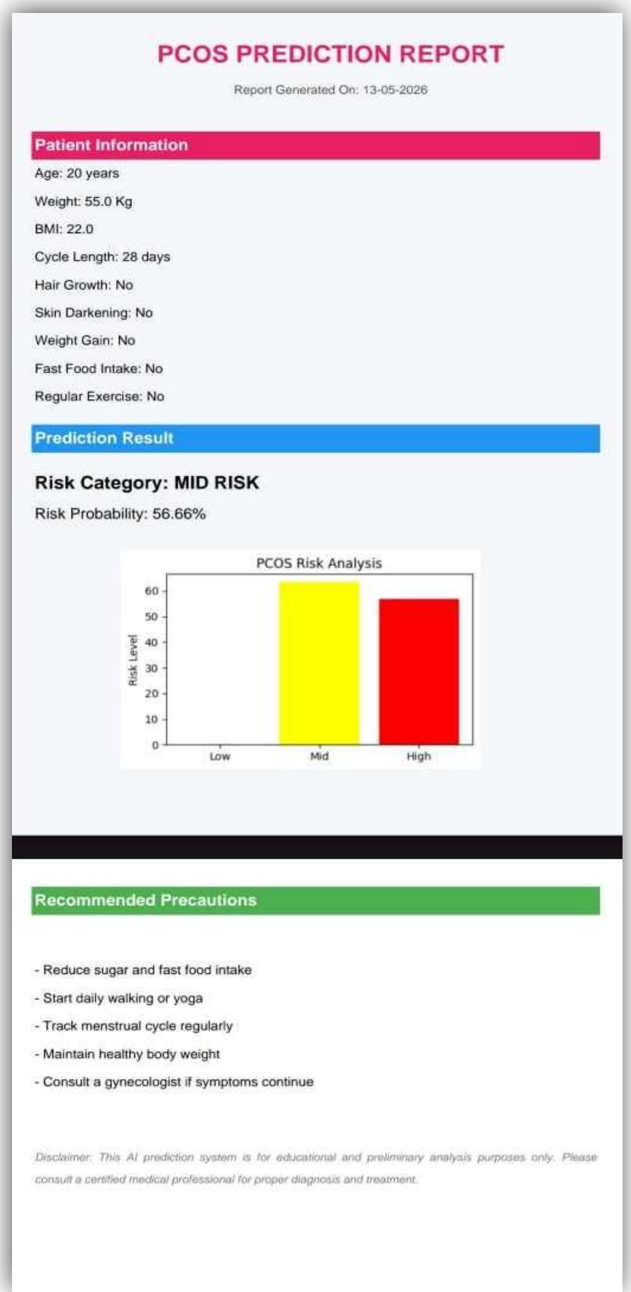

Fig 20: Mid-Risk Prediction Report

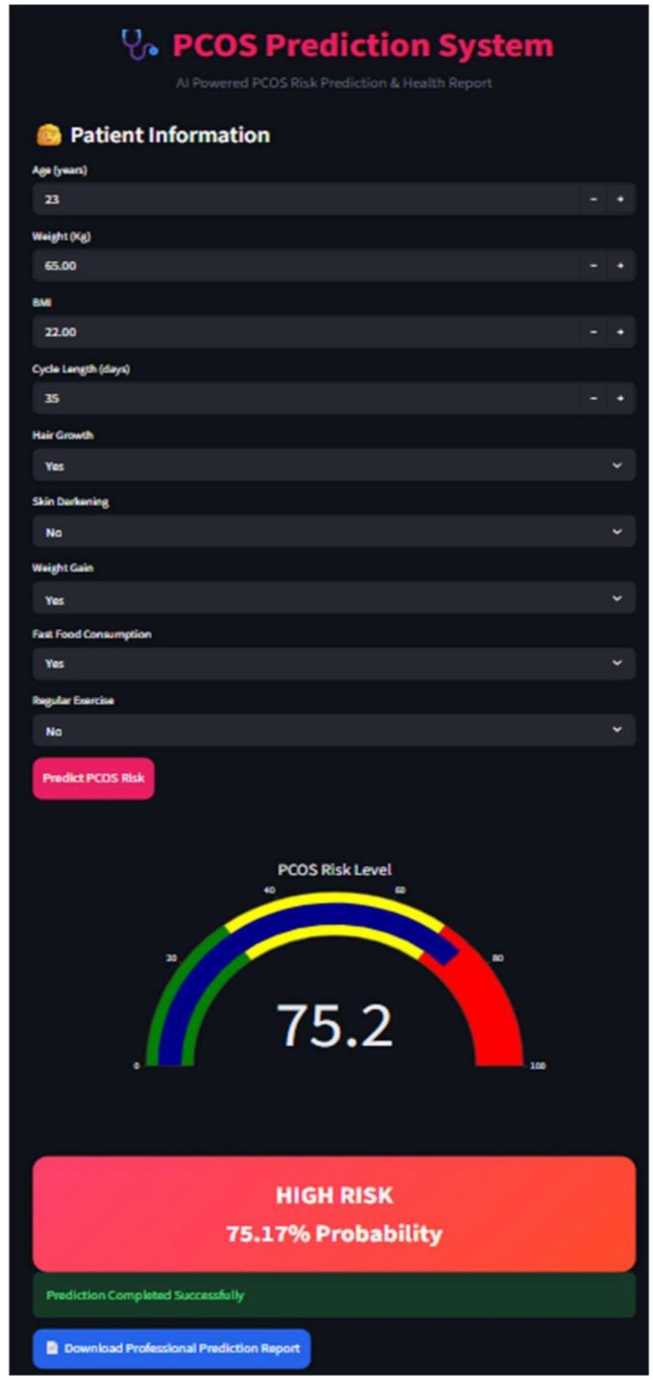

Fig 21: High Risk Prediction Table

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Fig 22: High Risk Prediction Report

CONCLUSION

This research successfully demonstrates the implementation of an AI-driven predictive screening and reporting framework for polycystic ovary syndrome (PCOS). By conducting a rigorous evaluation across multiple gradient boosting architectures, the CatBoost classifier was established as the most effective core engine, delivering a superior testing accuracy of 91% and a diagnostic ROC-AUC of 0.95. This strong discriminatory capability ensures highly dependable risk assessments, directly mitigating the critical clinical challenge of false negatives.

The technical deployment via an interactive Streamlit web application bridges the operational gap between complex computational logic and practical utility. By translating numerical risk vectors into intuitive, color-coded gauge charts and automating the compilation of structured, professional PDF health reports, the system provides a dual-benefit infrastructure. It significantly minimizes the preliminary diagnostic burden on overextended healthcare facilities while simultaneously empowering women with immediate, accessible, and data-driven insights regarding their hormonal and reproductive well-being.

REFERENCES

[1] Azziz, R., et al. (2016). "Polycystic ovary syndrome." *Nature Reviews Disease Primers*.

[2] Teede, H. J., et al. (2018). "Recommendations from the international evidence-based guideline for the assessment and management of PCOS." *Human Reproduction*.

[3] Moran, L. J., et al. (2010). "Metabolic syndrome in polycystic ovary syndrome: A systematic review." *Human Reproduction Update*.

[4] Wolf, W. M., et al. (2018). "Geographic distribution of polycystic ovary syndrome prevalence." *Int. J. Environ. Res. Public Health*.

[5] Wild, R. A., et al. (2010). "Long-term health consequences of polycystic ovary syndrome." *The Journal of Clinical Endocrinology & Metabolism*.

[6] Ravi, D., et al. (2017). "Deep learning for health informatics." *IEEE Journal of Biomedical and Health Informatics*.

[7] Chen, T., & Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." *ACM SIGKDD*.

[8] Denny, A., et al. (2019). "PCOS prediction using machine learning." *International Journal of Computer Applications*.

[9] Bharti, R., et al. (2020). "Prediction of Post-Traumatic Stress Disorder using Machine Learning." *Journal of Medical Systems*.

[10] Prokhorenkova, L., et al. (2018). "CatBoost: unbiased boosting with categorical features." *NeurIPS*.

[11] Streamlit Documentation (2024). "Streamlit: The fastest way to build data apps."

[12] Ke, G., et al. (2017). "LightGBM: A highly efficient gradient boosting decision tree." *NeurIPS*.

[13] Shrestha, R. B., et al. (2021). "Clinical Decision Support Systems in Medicine." *Health Informatics Journal*.

[14] Faust, O., et al. (2018). "Deep learning for healthcare: Applications and challenges." *Briefings in Bioinformatics*.

[15] Gerybaite, A., et al. (2022). "The role of AI in improving women's healthcare." *Women's Health Reports*.

[16] Lizneva, D., et al. (2016). "Criteria and phenotypes of polycystic ovary syndrome." *Fertility and Sterility*.

[17] Lundberg, S. M., & Lee, S. I. (2017). "A unified approach to interpreting model predictions." *NeurIPS*.

[18] Wang, F., et al. (2019). "The state of artificial intelligence in healthcare." *AI Magazine*.

[19] Mehrotra, P., et al. (2011). "Automated diagnosis of PCOS using machine learning." *IEEE BIBM*.

[20] Caruana, R., et al. (2015). "Intelligible models for healthcare." *ACM SIGKDD*.