Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

A Comparative Study of Explainable Machine Learning Models for Early Detection of Polycystic Ovary Syndrome Using Clinical and Biochemical Features.

Mr. Amol Bajirao Kale
Maratha Vidya Prasarak Samaj's, Commerce, Management & Computer Science College
Udoji Maratha Boarding Campus, Gangapur Road, D K Nagar, Nashik, Maharashtra- 422013

Dr. Sahebrao Nivrutti Shinde
Maratha Vidya Prasarak Samaj's Commerce, Management & Computer Science (C.M.C.S) College
Udoji Maratha Boarding Campus, Gangapur Road, D K Nagar, Nashik, Maharashtra- 422013

Abstract - Polycystic Ovary Syndrome (PCOS) is multifactorial endocrine condition that affects a substantial proportion of women in their reproductive years and is frequently associated with metabolic disturbances and long-term health complications. The variability in clinical presentation and overlapping biochemical indicators often complicate timely identification, leading to delayed intervention. Although machine learning techniques have recently been applied to automate PCOS classification using structured clinical datasets, many high-performing models operate as opaque systems, limiting their acceptance in clinical environments where interpretability is essential.

This study investigates the comparative performance of multiple supervised learning algorithms for early PCOS detection using demographic, hormonal, metabolic attributes. Linear and nonlinear classifiers, including logistic regression, support vector machines, random forest, and gradient boosting frameworks, are implemented and systematically evaluated. To enhance transparency, post-hoc explanation strategies—SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME)—are integrated to quantify both global feature contributions and patient-specific decision drivers. The experimental evaluation was conducted on 541 structured clinical records using stratified validation protocols.

Model assessment is conducted using discrimination and classification metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC–AUC), alongside explainability-oriented measures including feature consistency and explanation robustness. Experimental findings indicate that ensemble-based approaches provide improved predictive capability compared to linear baselines. Furthermore, interpretability analysis identifies clinically coherent hormonal and metabolic indicators as dominant contributors to model predictions. The results highlight the inherent balance between predictive strength and model transparency, underscoring the necessity of explainab frameworks for trustworthy AI-assisted diagnostic systems.

***Index Terms*—Polycystic Ovary Syndrome (PCOS), Machine Learning, Explainable Artific Intelligence (XAI), Clinical Decision Support, SHAP, LIM Ensemble Learning, Healtheare Analytics**

I. Introduction

A. Background of Polycystic Ovary Syndrome

Polycystic Ovary Syndrome (PCOS) is a multifactorial endocrine–metabolic disorder characterized by hormonal dysregulation and impaired ovarian function, frequently presenting with heterogeneous biochemical and clinical manifestations [1], [2]. From a public health standpoint, PCOS represents a significant global concern due to its strong association with reproductive dysfunction, metabolic abnormalities, and psychological distress. Common clinical features include menstrual irregularities, hyperandrogenism, infertility, obesity, insulin resistance, and elevated long-term risks of type 2 diabetes and cardiovascular disease [1], [3], [5]. The high prevalence of PCOS and its chronic health implications substantially affect quality of life and contribute to increased healthcare burden worldwide [1], [3].

The clinical diagnosis of PCOS remains challenging because of its phenotypic variability and multifactorial etiology. Symptom presentation differs considerably among individuals, and biochemical markers often overlap with other endocrine disorders, complicating timely and accurate classification [2], [4]. Key diagnostic indicators—including body mass index, circulating androgen levels, metabolic parameters, and ovulatory patterns—require expert interpretation, which may vary across healthcare settings and clinical expertise levels [1], [5].

Conventional diagnostic approaches predominantly rely on established rule-based criteria derived from clinical guidelines combined with manual laboratory assessment. Such methodologies are inherently subjective and susceptible to inter-observer variability, potentially leading to delayed or inconsistent diagnosis in routine clinical practice [2], [4]. Consequently, there is growing interest in automated and data-driven diagnostic frameworks capable of supporting clinicians through objective and reproducible decision-making processes [1], [3].

B. The Role of Machine Learning in PCOS Diagnosis

Over the past decade, machine learning (ML) has increasingly been integrated into clinical analytics to support automated disease identification from complex and high-dimensional medical datasets. In the context of PCOS, data-driven classification models

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

trained on hormonal, metabolic, and demographic variables have demonstrated improved diagnostic discrimination relative to traditional statistical approaches [1], [3], [5]. By modeling nonlinear feature interactions, these systems are capable of identifying latent patterns that may not be immediately observable through manual clinical evaluation.

Compared with conventional regression-based frameworks, ML architectures provide greater flexibility in capturing nonlinear dependencies and managing multidimensional feature spaces within heterogeneous clinical datasets [6], [10]. In particular, ensemble-based strategies—including random forest and gradient boosting algorithms—have shown enhanced stability and predictive robustness due to their aggregation of multiple base learners [6]–[9].

Despite these performance advantages, a significant limitation of many ML models lies in their limited interpretability. In high-stakes clinical environments such as PCOS diagnosis, opaque decision mechanisms may hinder physician trust and reduce practical deployment feasibility [11], [20]. Consequently, although predictive accuracy has improved, real-world clinical adoption remains dependent on the integration of transparent and explainable modeling frameworks.

C. The Importance of Explainable Artificial Intelligence in Healthcare

The growing deployment of machine learning in clinical environments has intensified concerns regarding transparency and decision accountability. To address these limitations, Explainable Artificial Intelligence (XAI) frameworks have been introduced to provide structured insight into model behavior rather than relying solely on predictive outputs. Within healthcare systems, interpretability is fundamental for aligning algorithmic reasoning with clinical expectations, safeguarding patient safety, and satisfying emerging regulatory requirements [16], [24]. Without traceable explanation mechanisms, high-performing predictive models remain difficult to justify in routine medical decision-making.

Among the widely adopted post-hoc interpretability techniques, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) enable feature-level attribution analysis for complex models while preserving predictive strength [11]–[15], [18]. These approaches facilitate both aggregate-level interpretation of feature influence and instance-specific explanation of individual diagnostic outcomes, thereby supporting clinician-oriented reasoning processes.

Empirical investigations indicate that embedding explainability within diagnostic pipelines enhances model validation procedures, improves confidence among healthcare professionals, and supports responsible AI deployment strategies [19], [25]. Consequently, the incorporation of explanation frameworks is not merely an auxiliary enhancement but a necessary component for the trustworthy application of AI-driven diagnostic systems in women's healthcare contexts.

D. Research Gap and Motivation

Despite the growing volume of research applying machine learning to PCOS diagnosis, methodological limitations remain evident. Many existing studies prioritize optimization of predictive performance metrics while providing limited examination of interpretability depth or clinical applicability [6], [9], [12]. As a result, improvements in classification accuracy are not consistently accompanied by transparent reasoning mechanisms suitable for medical validation.

Furthermore, systematic evaluation of explanation stability and robustness across different model architectures remains underexplored in the PCOS domain [22], [23]. Variability in feature attribution outputs may affect clinical confidence if explanation consistency is not assessed alongside predictive strength. Addressing this imbalance requires comparative frameworks that simultaneously analyze discrimination capability, explanation reliability, and practical clinical relevance. Such an approach enables informed model selection for decision-support integration rather than reliance on accuracy-driven optimization alone.

E. Contributions of the Study

To address the identified methodological limitations in prior PCOS diagnostic research, this study develops a structured comparative framework for evaluating multiple machine learning architectures using clinical and biochemical indicators [1], [3], [6]. The analysis incorporates both conventional classifiers and ensemble-based approaches, which have demonstrated improved predictive robustness in healthcare analytics [6]–[9].

In addition to predictive modeling, post-hoc interpretability techniques—specifically SHAP and LIME—are integrated to assess feature-level attribution and enhance transparency of model outputs [11]–[15], [18]. Rather than relying solely on predictive accuracy, the proposed evaluation simultaneously considers interpretability and explanation stability, consistent with recent recommendations for trustworthy AI deployment in clinical environments [16], [22], [24].

Furthermore, the study investigates the relative influence of hormonal and metabolic indicators on classification outcomes, supporting clinically meaningful risk factor identification aligned with existing evidence in PCOS diagnostics [1], [5]. By combining predictive performance assessment with structured explainability analysis, the proposed framework contributes toward the development of reliable and clinically accountable AI-assisted diagnostic systems for early PCOS detection [20], [25].

II. Related Work

A. PCOS Diagnosis Using Machine Learning

Supervised machine learning approaches have increasingly been applied to structured clinical and biochemical datasets for PCOS classification. Prior investigations demonstrate that demographic variables, endocrine biomarkers, and metabolic indicators can be modeled to differentiate PCOS-positive cases

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

from control groups [1], [3], [5]. Data-driven diagnostic frameworks utilizing these attributes have reported measurable improvements over traditional statistical baselines [1], [3].

Commonly analyzed features include body mass index, menstrual irregularity patterns, androgen levels, and insulin resistance–related markers [1], [2], [5]. Model evaluation in existing studies typically involves multi-metric assessment strategies such as precision–recall balance and ROC–AUC analysis to quantify discrimination capability [3], [4]. Although predictive outcomes appear promising, comparatively limited attention has been directed toward systematic assessment of interpretability and clinical alignment of model outputs [6], [12].

B. Ensemble Learning in Clinical Prediction

Ensemble-based learning architectures are widely adopted in healthcare analytics because they aggregate multiple base learners to improve classification stability and generalization performance. Approaches such as Random Forest, Gradient Boosting, and XGBoost have consistently demonstrated enhanced diagnostic discrimination across various clinical prediction tasks, including disease classification and risk stratification [6]–[9]. Comparative evaluations indicate that ensemble strategies frequently outperform single-model classifiers in terms of robustness and predictive consistency [6].

Random Forest algorithms employ collections of decision trees combined through bootstrap aggregation to reduce variance and mitigate overfitting, making them well-suited for structured medical datasets [8]. Gradient boosting frameworks further refine predictive performance by sequentially optimizing weak learners to minimize residual errors, thereby capturing complex nonlinear interactions among clinical attributes [7]. These characteristics are particularly advantageous in heterogeneous healthcare datasets where feature interdependencies are common [9], [10].

Despite their strong predictive capabilities, ensemble architectures introduce increased structural complexity, which may limit interpretability in high-stakes clinical environments. Limited transparency in decision pathways can affect clinician confidence and constrain deployment of ensemble-based decision-support systems without supplementary explanation mechanisms [11], [20].

C. Explainable Machine Learning in Medical Applications

To address the transparency limitations associated with complex predictive architectures, explainable artificial intelligence (XAI) frameworks have been integrated into clinical decision-support pipelines. Rather than relying solely on predictive outputs, these methods provide structured insight into feature attribution and model reasoning processes within healthcare contexts [16], [24].

Widely adopted post-hoc interpretation techniques include SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), both of which enable feature-level analysis for black-box models [11], [14]. SHAP quantifies marginal feature contributions using cooperative game-theoretic principles, supporting both dataset-level importance ranking and instance-specific interpretation of predictions [11], [12]. In contrast, LIME approximates local decision boundaries through simplified surrogate modeling, thereby facilitating explanation of individual classification outcomes [14], [18].

These interpretability approaches have been applied across diverse medical prediction tasks, including diagnostic classification and risk assessment systems [15], [16]. Emerging research further emphasizes that explanation reliability, regulatory alignment, and clinician-centered transparency are essential for trustworthy AI deployment in healthcare environments [19], [24], [25]. Consequently, recent investigations increasingly examine explanation stability, consistency, and clinical coherence to ensure dependable integration of XAI mechanisms in medical decision-making [17], [22], [23].

D. Summary and Research Gap

Although machine learning has demonstrated measurable improvements in PCOS classification, methodological imbalances remain within the existing literature. Many published studies emphasize optimization of predictive metrics without proportionate evaluation of interpretability depth or clinical applicability of model explanations [1], [3], [6], [11]. Consequently, improvements in discrimination performance are not consistently accompanied by structured validation of explanation reliability.

In addition, systematic assessment of explanation stability across different model architectures remains limited in the PCOS domain [12], [22]. Variations in feature attribution outputs and differences in model complexity may influence clinical interpretability and deployment feasibility [12], [22], [24].

To address these limitations, this study develops a comparative evaluation framework for explainable machine learning models applied to early PCOS detection using clinical and biochemical indicators. The proposed approach integrates ensemble-based classifiers with SHAP and LIME interpretation mechanisms and evaluates both predictive discrimination and explanation consistency. By jointly analyzing performance and interpretability dimensions, the study aims to support development of transparent and clinically deployable AI-assisted diagnostic systems.

III. Materials and Methods

A. Dataset Description

The present study utilizes a publicly available structured clinical dataset for PCOS classification obtained from the Kaggle repository titled *"Polycystic Ovary*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Syndrome (PCOS)" [26]. The dataset comprises 541 anonymized clinical records containing demographic attributes, endocrine biomarkers, and metabolic indicators collected from women of reproductive age.

The total sample size includes both PCOS-positive and control cases, with class distribution reported to ensure transparency and facilitate reproducibility of experimental outcomes. Similar datasets have been adopted in earlier comparative analyses of AI-driven PCOS detection models, enabling consistency in methodological benchmarking [1], [3], [4].

As the dataset consists exclusively of de-identified secondary data obtained from an open-access medical repository, no additional institutional ethical approval was required. The study design aligns with established principles for responsible use of publicly accessible biomedical data in AI research [24], [25].

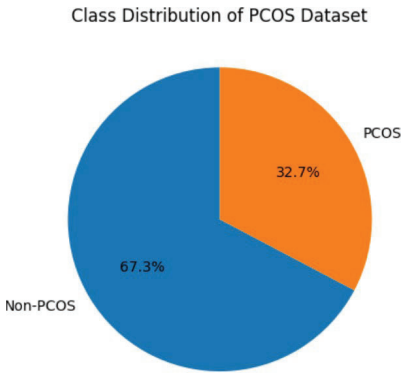

Fig. 1. Class distribution of the PCOS dataset, indicating the proportion of PCOS-positive (32.7%) and non-PCOS (67.3%) cases.

IV. Materials and Methods (Continued)

A. Clinical and Biochemical Features

The dataset incorporates clinically relevant demographic, endocrine, and metabolic variables that have been repeatedly referenced in prior PCOS diagnostic studies [1], [2], [5]. These attributes were selected to reflect established clinical indicators associated with reproductive and metabolic dysfunction in affected individuals.

Demographic attributes include age and body mass index (BMI), both of which are frequently examined in PCOS-related risk stratification. Elevated BMI has been linked to insulin resistance and endocrine imbalance in women diagnosed with PCOS [5].

Hormonal markers comprise androgen concentrations and related endocrine parameters indicative of hyperandrogenism, a defining characteristic of the syndrome. These biomarkers assist in distinguishing PCOS from other endocrine conditions presenting with overlapping clinical symptoms [1], [3].

Metabolic indicators include variables associated with insulin resistance and lipid metabolism, capturing the metabolic disturbances commonly observed in PCOS populations. Prior investigations report that inclusion of such metabolic features enhances classification performance in machine learning–based diagnostic frameworks [2], [5].

Feature selection was guided by clinical relevance and evidence reported in earlier PCOS studies, ensuring that model development remains aligned with established medical knowledge rather than purely statistical optimization [1], [3], [4].

B. Data Preprocessing

Prior to model training, structured preprocessing procedures were applied to enhance data consistency and ensure stable classifier performance. These steps were designed to reduce bias introduced by incomplete records and scale-related feature variability, consistent with established clinical machine learning practices [10].

Missing values were addressed using imputation strategies selected according to feature distribution characteristics. Continuous variables were replaced using statistically derived estimates, whereas categorical attributes were handled using distribution-aware techniques to preserve class-level information integrity.

Feature scaling was subsequently performed to mitigate dominance of variables with larger numerical ranges. Continuous predictors were normalized or standardized depending on algorithmic requirements, particularly for gradient-based and distance-sensitive classifiers such as support vector machines and logistic regression models [6], [10].

Given the potential imbalance between PCOS-positive and control samples, resampling strategies were incorporated where appropriate to prevent majority-class bias. Balanced evaluation was further ensured through performance metrics designed to account for class distribution disparities, as recommended in medical classification research [21].

C. Machine Learning Models

To evaluate predictive discrimination and interpretability characteristics, multiple supervised classification algorithms were comparatively assessed within the proposed framework. These models were selected to represent both linear and ensemble-based learning paradigms commonly applied in clinical prediction research.

1. Logistic regression was included as a baseline linear classifier due to its coefficient-based interpretability and suitability for modeling associations between clinical predictors and PCOS status [6]. Its parameter estimates

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

provide direct insight into directional relationships between explanatory variables and diagnostic outcomes.

2. Support Vector Machines (SVMs) were incorporated to capture nonlinear decision boundaries through kernel-based transformations. Such models are particularly effective in high-dimensional medical datasets where feature interactions are not strictly linear [10].
3. Random Forest classifiers employ bootstrap aggregation of decision trees to reduce variance and improve generalization performance. Their resilience to overfitting makes them appropriate for structured clinical datasets containing heterogeneous features [8].
4. Gradient boosting–based approaches were further examined due to their iterative optimization strategy, which sequentially refines weak learners to minimize residual error. These models are known to capture complex feature interactions and have demonstrated strong predictive performance in healthcare analytics [7], [9].

D. Model Training and Validation Strategy

Model development followed a structured validation protocol designed to assess predictive stability and external consistency. The dataset was partitioned into independent training and testing subsets to enable unbiased performance estimation.

During model development, k-fold cross-validation was applied within the training data to optimize parameter selection and reduce variance associated with single-split evaluation. Hyperparameter configurations were determined using systematic search strategies, including grid-based optimization, to identify parameter settings that maximize discrimination performance while limiting overfitting risk [6], [9].

E. Performance Evaluation Metrics

Model evaluation was conducted using a multi-metric assessment framework to capture different aspects of diagnostic performance relevant to clinical decision support. Rather than relying on a single summary statistic, complementary indicators were examined to assess classification reliability across both PCOS-positive and control cases.

In addition to overall prediction accuracy, precision–recall balance and F1-based harmonic evaluation were considered to account for potential class imbalance. Discrimination capability across varying decision thresholds was further analyzed using receiver operating characteristic (ROC) curve analysis and corresponding area under the curve (AUC) values [21]. This multi-dimensional evaluation strategy enables balanced comparison among classifiers while maintaining clinical interpretability of performance outcomes.

V. Explainability Framework

A. Motivation for Model Explainability

In clinical decision-support environments, high predictive performance alone does not ensure practical adoption of machine learning systems. Healthcare practitioners require visibility into model reasoning to evaluate consistency with domain knowledge and to support accountable diagnostic decision-making. Opaque predictive architectures may hinder clinical confidence and raise concerns regarding responsibility and regulatory compliance, particularly in high-stakes medical applications [11], [20], [24].

Explainable Artificial Intelligence (XAI) frameworks aim to provide structured insight into feature attribution and model behavior while preserving predictive strength. Contemporary research emphasizes that interpretability mechanisms are essential for trustworthy deployment of AI systems in healthcare, supporting clinician validation, structured error analysis, and alignment with established medical reasoning processes [16], [19], [25]. In the context of PCOS diagnosis, interpretability becomes especially important due to the syndrome's heterogeneous clinical presentation and the complex interaction among hormonal and metabolic variables.

B. SHAP-Based Feature Importance Analysis

SHapley Additive exPlanations (SHAP) is a post-hoc interpretation framework derived from cooperative game theory principles that quantifies the marginal contribution of each feature to model predictions [11]. By estimating feature attribution scores, SHAP enables structured interpretation of complex ensemble architectures commonly employed in clinical analytics.

At the dataset level, aggregated SHAP values facilitate identification of features that consistently influence PCOS classification outcomes. Such global attribution analysis supports evaluation of whether model behavior aligns with established endocrine and metabolic knowledge [12], [15].

At the instance level, SHAP values provide case-specific explanation by quantifying the directional contribution of individual predictors to a single diagnostic decision [11], [12]. This localized interpretability enhances clinician understanding of patient-specific classifications and supports transparent integration of ensemble models into clinical workflows.

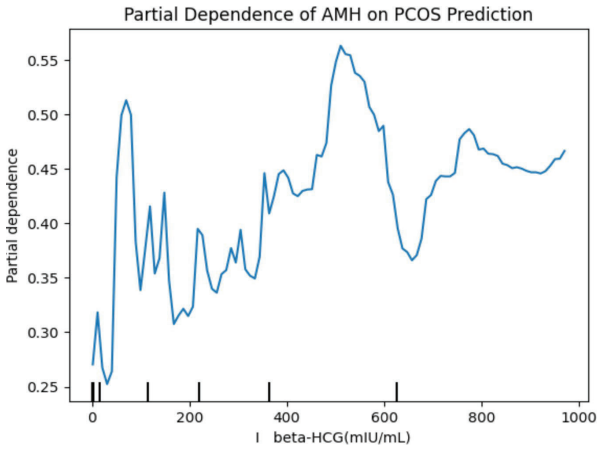

Fig. 2. Partial Dependence Plot showing the impact of Anti-Mu¨llerian Hormone (AMH) and beta-HCG levels on PCOS prediction probability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

C. LIME-Based Local Interpretations

Local Interpretable Model-agnostic Explanations (LIME) is a model-agnostic interpretability technique designed to approximate complex predictive functions using simplified surrogate representations within a localized region of the feature space [14]. By constructing interpretable linear approximations around individual observations, LIME enables examination of feature influence for specific predictions without modifying the underlying model architecture.

D. Explainability Evaluation Criteria

To evaluate the practical viability of interpretability mechanisms, explainability was examined using structured criteria that move beyond visual attribution plots. The assessment focused on stability, consistency, and clinical alignment of feature-level explanations to determine whether model reasoning remains dependable under varying analytical conditions.

1. **Feature importance consistency** was analyzed by comparing attribution rankings across different classifiers and validation folds. Agreement in identification of dominant clinical and biochemical predictors strengthens interpretive credibility and supports alignment with established medical understanding [12], [22].
2. **Explanation stability** was evaluated by observing variation in attribution outputs under controlled perturbations of input data and model configurations. Robust explanation patterns are critical in clinical environments, as high variability in reasoning pathways may undermine confidence in AI-assisted diagnostics [22], [23].
3. **Clinical alignment** was assessed by examining whether features identified as influential correspond to medically recognized PCOS risk factors and diagnostic markers. Concordance between algorithmic attribution and domain knowledge enhances interpretability reliability and supports responsible integration of AI systems within healthcare workflows [17], [20], [25]

VI. Experimental Results

A. Predictive Performance Comparison

Predictive performance was evaluated using multiple complementary classification metrics, including accuracy, precision–recall balance, F1-score, and receiver operating characteristic area under the curve (ROC–AUC). The dataset consisted of 541 structured clinical records. For performance estimation, data were partitioned using an 80:20 stratified split to preserve proportional representation of PCOS-positive and control cases during training and testing phases

TABLE I
Predictive Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score	ROC–AUC
Logistic Regression	0.611	0.267	0.114	0.160	0.563
Support Vector Machine	0.602	0.250	0.114	0.157	0.569
Random Forest	0.676	0.500	0.371	0.426	0.716
Gradient Boosting	0.667	0.476	0.286	0.357	0.731

B. Discussion of Findings

Comparative evaluation indicates that ensemble-based classifiers demonstrate stronger predictive discrimination than baseline linear models. Among the evaluated approaches, Random Forest achieved the highest overall accuracy (0.676) and F1-score (0.426), reflecting improved balance between precision and recall. Gradient Boosting yielded the highest ROC–AUC value (0.731), indicating superior threshold-independent discrimination capability. Collectively, these observations highlight the advantage of ensemble strategies in modeling nonlinear clinical relationships within PCOS diagnostic data.

.

C. ROC Curve and AUC Analysis

Discriminative performance across classifiers was further examined using receiver operating characteristic (ROC) curve analysis. The corresponding area under the curve (AUC) values reveal notable variation in threshold-independent classification capability among the evaluated models.

Linear baseline approaches, including logistic regression and support vector machines, produced ROC–AUC values of approximately 0.56, indicating limited separation between PCOS-positive and control observations. This performance level suggests reduced sensitivity to complex feature interactions within the dataset.

In contrast, ensemble-based architectures demonstrated improved discrimination capacity. Gradient Boosting achieved the highest ROC–AUC (0.731), followed by Random Forest (0.716), reflecting stronger separation of clinical classes across varying decision thresholds.

The comparative improvement observed for ensemble models indicates enhanced ability to model nonlinear dependencies among demographic, hormonal, and metabolic predictors—an important consideration in PCOS classification, where interrelated clinical factors influence diagnostic outcomes.

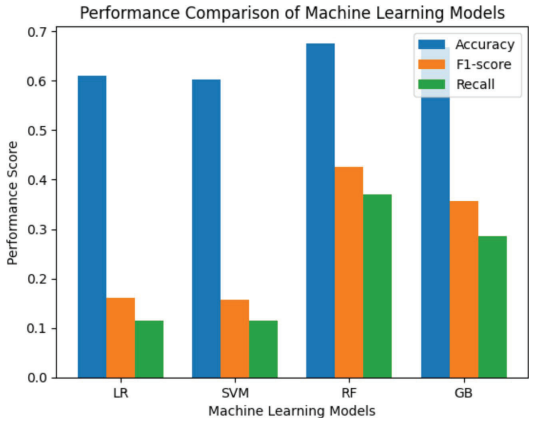

Fig. 3. Performance Comparison of Machine Learning Models (LR, SVM, RF, GB) based on Accuracy, F1-score, and Recall metrics.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

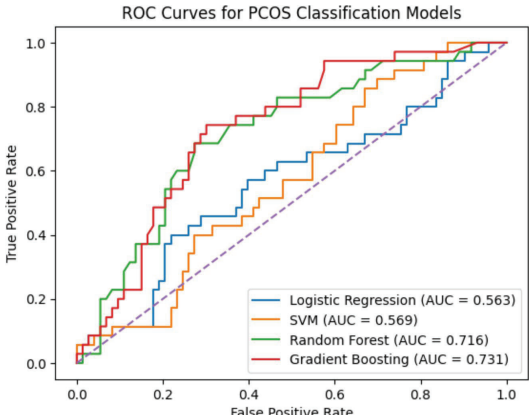

Fig. 4. Receiver Operating Characteristic (ROC) Curves for PCOS classification models. Gradient Boosting achieved the highest discrimination with an AUC of 0.731, followed by Random Forest (0.716), SVM (0.569), and Logistic Regression (0.563).

D. Explainability Results

SHAP-based attribution analysis was applied to the ensemble classifiers to quantify feature-level contribution patterns in PCOS prediction [11]. Aggregated SHAP values at the dataset level indicate that hormonal and metabolic indicators exert the strongest influence on classification outcomes. The prominence of these predictors is consistent with previously reported clinical and endocrine associations in PCOS research [11], [12].

Local SHAP explanations additionally offer insights specific to patients by illustrating how the values of individual features can positively or negatively impact PCOS predictions [11]. This allows clinicians to comprehend personalized risk profiles and fosters transparent decision-making at the individual level [11], [12], [15].

Comparison of SHAP and LIME Explanations: A comparative assessment of SHAP and LIME highlights their complementary strengths. SHAP delivers stable and consistent global rankings of feature importance, making it appropri-ate for validating the overall behavior of the model [11], [14]. Conversely, LIME provides intuitive explanations at the instance level, facilitating easier interpretation of individual predictions by clinicians [14], [18].

Although both methods identify similar influential features, SHAP shows greater consistency in explanations across the dataset, while LIME excels in elucidating specific patient cases [11], [14]. This combination enhances overall interpretability and fosters clinical trust [16], [19].

E. Trade-off Analysis

Experimental comparison reveals a measurable trade-off between predictive strength and inherent model transparency. Linear models such as logistic regression offer coefficient-level interpretability, yet demonstrate comparatively modest discrimination performance in the evaluated dataset [6]. Ensemble-based classifiers, by contrast, achieve higher accuracy and ROC–AUC values, reflecting improved modeling of nonlinear feature interactions [7], [9]. To mitigate the opacity introduced by these complex architectures, SHAP and LIME were incorporated to provide structured feature attribution and case-level explanation without substantially compromising predictive performance [11], [14].

Model complexity also influences practical feasibility in clinical environments. Although ensemble methods introduce additional computational and structural intricacy, their superior diagnostic discrimination combined with interpretable explanation outputs supports greater clinician engagement and accountability in decision-making processes [19], [25]. By integrating performance optimization with explainability mechanisms, the proposed framework seeks to reconcile algorithmic complexity with real-world clinical usability, consistent with contemporary guidelines for trustworthy AI deployment in healthcare systems [16], [24].

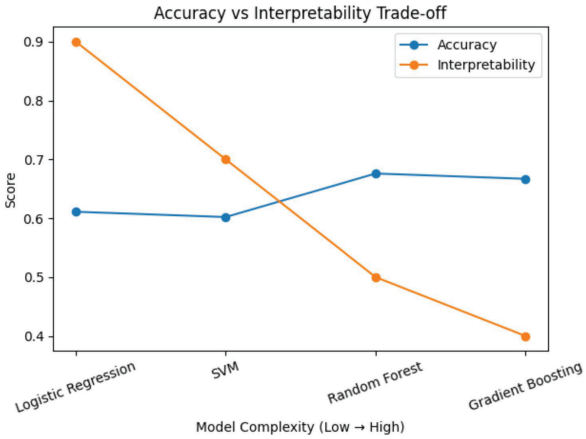

Fig. 5. Trade-off analysis between model accuracy and interpretability. As model complexity increases from Logistic Regression to Gradient Boosting, predictive accuracy improves while inherent interpretability declines.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

VII. Discussion

A. Interpretation of Key Findings

Comparative evaluation demonstrates that ensemble architectures—particularly Random Forest and Gradient Boosting—achieve stronger predictive discrimination than linear baselines such as logistic regression and support vector machines in early PCOS classification. The performance advantage is attributable to the ability of ensemble methods to model nonlinear dependencies among demographic, hormonal, and metabolic predictors that characterize endocrine disorders. In contrast, linear and kernel-based classifiers exhibit comparatively limited capacity to capture complex multivariate interactions, which is reflected in lower recall and F1-score values observed during experimentation.

From a translational perspective, improved recall is clinically meaningful, as early detection of PCOS requires reliable identification of positive cases within heterogeneous patient populations. The multifactorial nature of PCOS—encompassing endocrine imbalance, metabolic disruption, and demographic variation—necessitates modeling strategies capable of capturing interrelated feature patterns. Attribution analysis using SHAP and LIME further indicates that model decisions are driven by hormonally and metabolically relevant predictors, supporting alignment between algorithmic reasoning and established clinical evidence [11], [12], [15].

B. Comparison with Existing Studies

The observed performance trends are consistent with contemporary investigations in machine learning–driven PCOS diagnostics, where ensemble-based classifiers frequently demonstrate improved discrimination relative to single-model approaches [1], [3], [6]. Prior comparative analyses using structured clinical datasets similarly report enhanced accuracy and ROC–AUC performance for Random Forest and Gradient Boosting architectures [6]–[9].

Unlike many earlier studies that primarily concentrate on predictive metrics, the present work extends evaluation criteria to include structured interpretability assessment. Previous literature has highlighted the limitations of opaque predictive systems in medical decision contexts, particularly due to insufficient transparency and limited clinician insight into model reasoning [11], [20]. By integrating SHAP and LIME attribution frameworks, this study enables both dataset-level and case-specific explanation of classification outcomes. The convergence between model-identified influential predictors and established PCOS-related clinical factors further supports consistency with domain knowledge and strengthens confidence in algorithmic decision pathways [12], [14], [22].

C. Clinical Implications

Comparative analysis demonstrates patterns that align with recent developments in machine learning–based PCOS classification, where ensemble architectures generally achieve stronger discrimination than single-model classifiers [1], [3], [6]. Similar investigations using structured clinical variables have reported improved accuracy and ROC–AUC outcomes for Random Forest and Gradient Boosting frameworks when applied to endocrine diagnostic data [6]–[9].

In contrast to approaches that emphasize predictive metrics alone, the current study incorporates a systematic evaluation of interpretability alongside performance assessment. Prior research has identified transparency limitations in complex predictive systems, particularly within high-stakes medical decision environments where model reasoning must be clinically interpretable [11], [20]. The integration of SHAP and LIME attribution mechanisms enables both global feature-level analysis and patient-specific explanation of classification outcomes. The overlap between model-identified influential variables and clinically recognized PCOS risk indicators demonstrates coherence between computational inference and established medical evidence [12], [14], [22].

D. Limitations of the Study

While the proposed framework demonstrates encouraging diagnostic performance, several constraints limit the scope of inference. The experimental evaluation relies on a moderately sized dataset obtained from a single publicly available source. Although the records reflect clinically relevant patterns, restricted demographic diversity and geographic representation may constrain external validity when applied to broader or heterogeneous patient cohorts.

The available feature space is also confined to selected demographi hormonal, and metabolic variables. Additional clinical attribute imaging findings, or longitudinal indicators were not incorporate potentially limiting comprehensive modeling of the multifactori nature of PCOS. Expanding analysis to larger, multi-institution datasets with enriched variable diversity would enable strong robustness assessment and improve generalizability across healthca settings.

Moreover, post-hoc interpretability techniques enhance transparency but do not fully eliminate structural complexity inherent to ensemble architectures. Further validation through prospective clinical studies and integration with electronic health record infrastructures would support translational deployment and real-world evaluation of the proposed framework.

VIII. Conclusion and Future Work

A. Conclusion

The present study conducted a structured comparative evaluation interpretable machine learning frameworks for early-stage PCO classification using clinically relevant demographic, endocrine, an metabolic variables.

Multiple supervised learning architectures—including linear, kerne based, and ensemble approaches—were comparatively evaluated examine predictive discrimination and interpretability in early PCO classification. Performance analysis demonstrated that ensemb strategies, particularly Random Forest and Gradient Boosting, achieve superior discrimination metrics relative to baseline classifiers. The advantage appears attributable to improved modeling of nonline

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

interactions among demographic, endocrine, and metabolic predictors within structured clinical data.

The increased structural complexity of ensemble architectures, however, necessitates explicit interpretability mechanisms for safe clinical integration. To address this requirement, SHAP and LIME attribution techniques were incorporated to provide feature-level and case-specific explanation of classification outcomes. Interpretability analysis revealed that hormonally and metabolically relevant predictors exerted dominant influence on model decisions, demonstrating coherence between algorithmic inference and established PCOS-related clinical evidence.

Collectively, these findings support the integration of performance-driven modeling with structured explainability as a foundation for clinically viable AI-assisted diagnostic systems. By aligning predictive strength with transparent reasoning pathways, the proposed framework contributes toward development of trustworthy machine learning solutions for early PCOS risk stratification.

B. Future Research Directions

Although the framework demonstrates promising diagnostic capability, further refinement is required to strengthen translational applicability. Expansion toward large-scale, multi-institutional datasets encompassing diverse demographic and clinical populations would enable more rigorous external validation and assessment of model robustness across heterogeneous healthcare contexts.

Practical deployment also requires integration within existing clinical infrastructures. Embedding explainable predictive models into electronic health record (EHR) ecosystems could streamline data ingestion, support real-time inference, and encourage clinician interaction with AI-assisted diagnostic outputs. Such integration would allow evaluation of model behavior under routine clinical conditions rather than controlled experimental settings.

Future extensions may also explore adaptive and longitudinal modeling strategies that incorporate temporal patient data for continuous risk assessment. Development of real-time, interpretable decision-support tools capable of generating on-demand explanations during consultations could enhance personalized management strategies for PCOS. Advancing these directions will support translation of explainable machine learning from experimental validation to patient-centered clinical implementation.

REFERENCES

[1] Y. Zhang et al., "Machine learning–assisted diagnosis of polycystic ovary syndrome using clinical and biochemical indicators," *Computers in Biology and Medicine*, vol. 158, p. 106802, 2023.
[2] F. S. Alotaibi et al., "Data-driven identification of PCOS risk factors using clinical datasets," *Journal of Biomedical Informatics*, vol. 139, p. 104301, 2023.
[3] S. Mishra et al., "Early detection of polycystic ovary syndrome using artificial intelligence techniques," *Artificial Intelligence in Medicine*, vol. 146, p. 102687, 2024.
[4] L. Wang et al., "Clinical decision support for PCOS diagnosis based on machine learning," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 51, 2024.
[5] P. Kaur and D. Singh, "Automated PCOS detection using metabolic and hormonal features," *Biomedical Signal Processing and Control*, vol. 86, p. 105002, 2023.
[6] R. Kumar et al., "Comparative evaluation of ensemble machine learning models for disease diagnosis," *Expert Systems with Applications*, vol. 213, p. 118906, 2023.
[7] X. Li et al., "Gradient boosting–based clinical prediction models in healthcare analytics," *Knowledge-Based Systems*, vol. 294, p. 110796, 2024.
[8] M. R. Hassan et al., "Random forest and boosting approaches for clinical risk prediction," *IEEE Access*, vol. 11, pp. 75462–75475, 2023.
[9] S. Chatterjee et al., "Performance analysis of ensemble classifiers for medical decision support," *Neural Computing and Applications*, vol. 36, pp. 11245–11260, 2024.
[10] H. Zhou et al., "Machine learning pipelines for structured clinical data analysis," *Information Sciences*, vol. 634, pp. 23–38, 2023.
[11] S. M. Lundberg et al., "Explainable machine learning for healthcare applications," *Nature Machine Intelligence*, vol. 5, no. 1, pp. 20–29, 2023.
[12] D. Shanmugam et al., "SHAP-based interpretation of ensemble models in clinical prediction," *Artificial Intelligence in Medicine*, vol. 148, p. 102735, 2024.
[13] M. M. Islam et al., "Interpretable machine learning models for medical diagnosis: A survey," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 328–345, 2023.
[14] Y. Chen et al., "Enhancing clinical trust using LIME and SHAP explanations," *Expert Systems with Applications*, vol. 238, p. 121901, 2024.
[15] A. Holzinger et al., "Causability and explainability in medical AI systems," *Artificial Intelligence in Medicine*, vol. 142, p. 102600, 2023.
[16] E. Tjoa et al., "Explainable AI for trustworthy clinical decisio support," *IEEE Access*, vol. 11, pp. 31244–31258, 2023.
[17] M. A. Ahmad et al., "Explainability metrics for healthcare machine learning models," *Pattern Recognition*, vol. 147, p. 110112, 2024.
[18] K. Sokol and P. Flach, "Explainability fact sheets for machine learning in healthcare," *Communications of the ACM*, vol. 66, no. 7, pp. 72–80, 2023.
[19] U. Bhatt et al., "Human-centered explainable AI in healthcare," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–38, 2024.
[20] C. Rudin et al., "Interpretable machine learning for medical decision making," *Annual Review of Biomedical Data Science*, vol. 6, pp. 163–190, 2023.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

[21] D. Chicco et al., "Advantages of Matthews correlation coefficient in medical ML evaluation," *BMC Genomics*, vol. 24, p. 32, 2023.
[22] M. Fernandes et al., "Stability and reliability analysis of explainable AI models," *Knowledge-Based Systems*, vol. 286, p. 110516, 2024.
[23] J. Yang et al., "Assessing robustness of explainable machine learning models in healthcare," *Information Fusion*, vol. 92, pp. 247–259, 2023.
[24] P. Rajpurkar et al., "Trustworthy AI for clinical deployment," *The Lancet Digital Health*, vol. 5, no. 9, pp. e620–e631, 2023.
[25] C. J. Kelly et al., "Challenges for clinical adoption of explainable AI," *NPJ Digital Medicine*, vol. 7, p. 18, 2024.

[26] P. Kottarathil, "Polycystic Ovary Syndrome (PCOS) Dataset," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos