Received: 17 May, 2025, Revised: 14 June, 2025, Accepted: 13 June, 2025, Online: 26 June, 2025

DOI: https://doi.org/10.55708/js0406003

Comparative Analysis of Supervised Machine Learning Models for PCOS Prediction Using Clinical Data

Ranyah Taha*[,1], Huda Zain El Abdin [2], Tala Musleh [3]

[1] Computer Science Dept., Al-Iman School, Bahrain

[2] Faculty of Science and Technology, Computer Science Department, University of Middlesex, London, Hendon, United Kingdom

[3] Pharmacy Department, College of Health and Sports Sciences, University of Bahrain, Bahrain

Corresponding author: Huda Zain El Abdin, Wokingham, United Kingdom, hudaa.z@icloud.com , +447832032740

ABSTRACT: Polycystic Ovary Syndrome (PCOS) is a prevalent hormonal disorder affecting women of reproductive age, commonly resulting in irregular menstrual cycles, elevated androgen levels, and the presence of polycystic ovaries. It is a major cause of infertility and is often linked with metabolic complications such as insulin resistance and obesity. Symptoms vary and may include acne, excessive hair growth, weight gain, and hair thinning. Early detection and proper management through lifestyle interventions and medical treatment are crucial to mitigating long-term health risks. This study investigates the classification performance of seven supervised machine learning algorithms—Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Classifier (GBC), Adaptive Boosting (AdaBoost), and Multi-Layer Perceptron (MLP)—using clinical and lifestyle data related to PCOS. The models were evaluated using accuracy, precision, recall, F1 score, and ROC AUC metrics. LR consistently outperformed the other models, achieving the highest accuracy (91.7%), precision (96%), and Receiver Operating Characteristics -Area Under the Curve (ROC AUC) (96.8%), while also maintaining a strong balance in recall and F1 score. This outstanding performance is attributed to the linear nature of the dataset and the efficiency, simplicity, and generalizability of LR, making it particularly suitable for this classification task. This study introduces a novel approach for predicting PCOS by integrating advanced data preprocessing techniques with a focus on model simplicity and interpretability. The predictive performance of LR was further enhanced through the application of the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance and Analysis of Variance (ANOVA) F-score-based feature selection to identify the most statistically significant predictors. This approach not only achieved high predictive accuracy but also ensured transparency and ease of deployment, making it highly applicable for clinical decision-support systems aimed at early and accurate PCOS diagnosis.

KEYWORDS: Artificial Intelligence, Data Analysis, Polycystic Ovary Syndrome, Supervised Machine Learning, Medical Diagnosis.

1. Introduction

Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder that affects approximately 8–13% of women of reproductive age worldwide. Its prevalence, however, varies depending on ethnicity and diagnostic criteria. PCOS is characterized by hormonal imbalances, particularly elevated androgen levels, which manifest in symptoms such as irregular menstrual cycles, anovulation, and the presence of multiple ovarian follicles. These disruptions often lead to infertility and are commonly accompanied by metabolic complications, including obesity, insulin resistance, type 2 diabetes, and increased cardiovascular risk [1]. Despite its widespread occurrence and clinical implications, PCOS remains underdiagnosed due to its heterogeneous presentation and overlapping symptoms with other conditions. This diagnostic challenge underscores the need for advanced tools to enhance early detection and personalized care. Artificial intelligence (AI), particularly machine learning (ML), offers a promising solution by identifying complex, non-linear patterns within clinical and biochemical data—patterns that may be overlooked through conventional diagnostic approaches. Numerous studies have highlighted the potential of ML to augment clinical workflows in endocrinology, providing timely, data-driven support for healthcare professionals [1].

The presence of PCOS symptoms can vary significantly among women, it includes acne, extra body hair growth, hair thinning and obesity. As a matter of fact, symptoms differ from one person to another, which makes diagnosing PCOS challenging. Early and accurate diagnosis is essential for timely intervention to manage both reproductive and metabolic health risks [2].

The diagnostic process for PCOS remains challenging due to the diverse presence of symptoms across women. Typically, physicians rely on a combination of clinical assessments, blood tests, and pelvic ultrasound imaging. However, the absence of a comprehensive diagnostic tool also makes it hard to distinguish from other conditions leading to misdiagnosis or delays in diagnosis. Consequently, healthcare systems are seeking more advanced solutions to boost diagnostic accuracy to have efficient outcomes [3].

Advancements in AI, specifically in ML, have shown considerable promise in healthcare diagnostics. This is particularly for finding and classifying sophisticated diseases such as PCOS. For instance, supervised learning algorithms are demonstrating significant capability by uncovering hidden patterns within clinical and lifestyle data that was not readable by healthcare providers. The availability of electronic health records (EHRs) and patient data are rapidly increasing. Furthermore, AI-driven solutions could improve the prediction of PCOS diagnosis, enabling tailored and patient-specific treatments [4].

This study investigates the effectiveness of seven ML algorithms—LR, NB, SVM, RF, GBC, AdaBoost and MLP—in identifying PCOS using a dataset sourced from Kaggle. Following the CRISP-DM framework, the study applies a structured approach to data analysis and model development, incorporating patient data related to symptoms and lifestyle factors. The performance of each model is assessed using precision, recall, F1 score, and ROC AUC to enable a comparative evaluation of their strengths and limitations.

The findings aim to inform the development of AI-based diagnostic tools that support clinicians in diagnosing PCOS more accurately and efficiently, thereby enhancing clinical decision-making.

The study is structured into the following sections: literature review, methodology, data description and preprocessing, model implementation, results, discussion, conclusion, and future recommendations.

2. Literature Review

Several studies in recent years have used different ML techniques to diagnose and predict PCOS. Utilising clinical and physiological dataset to augment prediction accuracy. These approaches enhance distinct algorithms and data preprocessing methods for the aim of capturing patterns that assist in early and reliable PCOS detection.

In [5], the Decision Tree (DT), RF, and SVM algorithms were applied to a clinical dataset containing features such as Body Mass Index (BMI), insulin levels, and follicle count to predict the presence of PCOS. Among the models tested, the RF classifier achieved the highest accuracy of 89.5%. The study emphasized that ensemble models like RF are particularly effective in capturing complex relationships and interdependencies among clinical features.

Similarly, authors [6] used LR, NB, and KNN to analyse a dataset of 520 PCOS cases. In terms of model performance development, the study focused on feature selection techniques such as chi-square and recursive feature elimination. LR revealed strong predictive capability with an accuracy of 85.3%, especially when hormonal and metabolic attributes were emphasized. This demonstrates the strength of tree-based models in the clinical field.

In a more recent analysis, authors in [7] implemented DL models accompanied with traditional supervised classifiers on a refined clinical dataset. The study compared Artificial Neural Networks (ANN) with SVM, DT, and XGBoost. Despite the fact that ANN achieved the highest accuracy of 91.2%, the authors highlighted that simpler supervised model like XGBoost provided competitive results with lower computational costs, supporting their practicality for clinical integration.

In the imaging domain, researchers [8] proposed a model interpretability by combining DT classifiers with SHapley exPlanations (SHAP), a method that collaborates each independent feature to contribute to accurate predictions. This approach assembled the authors to generate a ranked list of features based on their impacts on the model's output. Nevertheless, testosterone levels and the luteinizing hormones (LH) to follicle-stimulating hormone (FSH) ratio emerged as dominant predictors lining up with clinical indicators of PCOS. Through the visualization of feature importance at both the population and patient-specific levels, the study provided a clearer understanding of the model's reasoning, which contributes to greater clinical confidence and interpretability in automated diagnostic applications.

Furthermore, authors [9] established a cloud-based diagnostic system trained on three different medical datasets taken from medical centres. AI algorithms analysed images, focusing on DNA content with cell nuclei. It validated the value of feature specificity such as DNA content as PCOS markers. Based on the results, these images were derived to a cloud-based platform for evaluation and assessments. Results achieved accuracy between 86% and 89%.

In addition, Arya [10] proposed a two-step approach to medical diagnosis that merges both supervised and unsupervised learning techniques. Starting with k-means clustering was used to group similar patient records. Followed by, analysing the clustered groups using supervised classification models, DT and SVM models to predict diagnosis. This combined method improved the accuracy of the system, reaching a prediction accuracy of 87.5%, and highlighted how blending ML techniques.

In the use of Graph Neural Network (GNN), Boll, et al. [11] acknowledges relationships between variables in EHRs. By treating clinical data as a network each variable is a node, and the connections reflect how these variables interact. As a result, patient information was modelled in a meaningful way. This graph-based approach achieved a strong AUC score of 89%, showing significant clinical prediction outcomes using advanced Deep Learning (DL) techniques.

Similarly, authors in [12] developed a Light Gradient Boosting Machine (LightGBM) model in conjunction with SHAP to identify and prioritise features relevant to PCOS diagnosis. The analysis highlighted the significance of anti-Müllerian hormone (AMH) levels and clinical signs such as hirsutism in prediction PCOS. As a result, the model achieved AUC of 93%, indicating high performance. In another notable comparison, Wang, et al. [13] implemented SVM, GBC, and MLP on PCOS datasets with categorical and numerical features. MLP achieved the highest F1 score 92%, demonstrating DL's ability to capture nonlinear relationships in diverse data formats. However, SVM maintained excellent generalization with less overfitting.

Additionally, authors in [14] examined the performance of LR, SVM, and MLP for early PCOS detection using lifestyle data (e.g., activity, sleep). Results show LR proved superior in AUC and interpretability, confirming its dominance in structured health data settings. Specifically, the study documented an AUC of 82.3% for the LR model, highlighting its robust performance.

Addressing the challenge of class imbalance, authors in [15] conducted an analysis on distinct algorithms, RF AdaBoost, and GBC on datasets with imbalanced PCOS class distributions. By applying SMOTE for balance, GBC performed best in handling rare class detection, with an AUC of 94.2%, followed closely by AdaBoost.

Similarly, authors in this study [16] developed predictive models using four ML methods: LR, SVM, GBC trees, and RF. It focused on hormone values (follicle-stimulating hormone, luteinizing hormone, oestradiol, and sex hormone-binding globulin) were combined to create a multilayer perceptron score using a neural network classifier. The models achieved AUC values of 85%, 81%, 80%, and 82%, respectively. Significant positive predictors of PCOS diagnosis across models included hormone levels and obesity; negative predictors included gravidity. The study illustrates the potential benefits of integrating AI tools into EHRs to facilitate earlier detection of PCOS.

Finally, researchers in [17] proposed three lightweight DL models LSTM-based, CNN-based, and CNN-LSTM-based for automated PCOS prediction. To address the imbalanced nature of the dataset, the SMOTE was employed. The models achieved accuracies of 92.04%, 96.59%, and 94.31%, with corresponding ROC-AUC values of 92.0%, 96.6%, and 94.3%. The study highlights the effectiveness of lightweight DL models in delivering high performance with fewer trainable parameters, making them suitable for resource-constrained environments.

Previous studies have utilized various ML algorithms to enhance PCOS diagnosis and prediction. Among these, RF demonstrated strong predictive capabilities by capturing complex, non-linear relationships, achieving accuracies up to 89.5%. LR was also widely used due to its simplicity and interpretability, particularly effective with structured clinical and lifestyle data, achieving accuracies above 85%.

SVM provided good generalization performance, especially on smaller datasets, but was sometimes outperformed by DL models on larger datasets. DL approaches, including ANN, CNN, and LSTM, achieved the highest accuracies, reaching up to 96.59% with CNN-LSTM architectures, though they required higher computational resources.

Tree-based ensemble models such as GBC and XGBoost delivered competitive results with lower computational costs, making them suitable for clinical environments. GBC particularly excelled in handling imbalanced datasets, achieving AUC values over 94%. Recently, advanced models like GNN were introduced to model complex relationships in electronic health records, achieving an AUC of 89%.

In summary, although DL models achieved the highest prediction accuracies, RF and GBC provided a balanced trade-off between performance, interpretability, and computational efficiency, making them highly applicable in practical clinical scenarios.

3. Research Methodology and approach

3.1. Background of the Research Study

This research was conducted using Google Colab as the primary development environment, with Scikit-learn as the main Python library for implementing ML models. A total of seven classification algorithms were employed to analyse and classify PCOS cases. The models used include LR, NB, SVM, RF, GB, AdaBoost, and MLP. Each algorithm was trained and evaluated to assess its effectiveness in accurately identifying PCOS based on clinical and lifestyle features.

The selection of these specific algorithms—LR, NB, SVM, RF, GBC, AdaBoost, and MLP—was driven by their complementary strengths in handling structured clinical data. LR offers high interpretability and computational efficiency, making it ideal for linear relationships within medical datasets. NB is well-suited for smaller datasets and performs effectively under the assumption of conditional feature independence. SVM is robust in high-dimensional spaces and generalizes well across complex boundaries. Ensemble methods such as RF, GBC, and AdaBoost are powerful in modeling non-linear interactions and addressing class imbalance, which are common in PCOS-related data. Lastly, MLP, a type of artificial neural network, was included for its ability to capture deep non-linear relationships. This diverse algorithm selection enables a comprehensive comparison across linear, probabilistic, ensemble-based, and neural learning paradigms, enhancing the model's applicability to the multifactorial nature of PCOS.

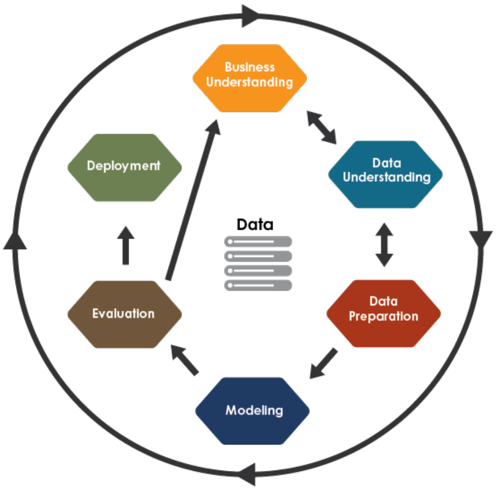

Figure 1: Phases of the CRISP-DM Methodology

The study followed the CRISP-DM methodology, a widely recognized framework for structuring ML projects. This approach consists of six key phases: defining project objectives (business understanding), exploring and analysing the dataset (data understanding), organizing and cleaning data for analysis (data preparation), developing and tuning ML models (modelling), assessing model performance (evaluation), and preparing the model for practical application (deployment) [18]. Adopting this structured workflow ensured clarity, consistency, and effectiveness throughout the project, ultimately contributing to the reliable and accurate results presented in Figure 1.

3.2. Dataset Description

The dataset used in this study was retrieved from Kaggle, a widely recognized platform for data science competitions and open-access datasets [19]. It contains clinical, biochemical, and lifestyle-related information collected from 541 female patients to support the prediction and diagnosis of PCOS. The dataset includes 44 features, including a binary target variable, PCOS (Y/N), where a value of 1 indicates a confirmed diagnosis of PCOS and 0 denotes its absence.

The features span several categories. Demographic and anthropometric variables include age, weight, height, BMI, and blood group. Vital signs such as pulse rate, respiratory rate, and blood pressure are included. Reproductive health indicators—like menstrual cycle regularity and pregnancy status—are complemented by hormonal measurements including AMH, FSH, LH, the FSH/LH ratio, and Beta-HCG. The dataset also captures symptoms and lifestyle factors, such as hair loss, acne, skin pigmentation, weight gain, hirsutism, fast food intake, and physical activity. Furthermore, ultrasound features detail follicle count and size in each ovary, along with endometrial thickness.

Notably, this dataset does not contain some of the core hormonal biomarkers typically used in the clinical diagnosis of PCOS, such as estrogen, progesterone, and testosterone. The absence of these indicators constitutes a key limitation of the dataset provided via Kaggle and was not a modeling decision but rather a constraint imposed by data availability. In real-world clinical practice, these hormones are fundamental to differential diagnosis and are often among the first parameters assessed alongside imaging. Their exclusion may restrict the model's ability to fully replicate the diagnostic reasoning employed by clinicians and can limit generalizability to broader patient populations. Future studies will aim to incorporate such hormonal data to enhance both predictive performance and clinical validity.

Additionally, the dataset does not include crucial demographic attributes such as ethnicity, geographical origin, and socioeconomic status—factors that significantly influence hormonal expression, symptomatology, and PCOS risk profiles. The lack of these variables introduces potential bias and restricts the fairness and applicability of the model across diverse populations. This limitation will be acknowledged explicitly in the revised manuscript, and future research will seek to mitigate these shortcomings through more inclusive and representative datasets. A summary of the dataset's attributes is provided in Table 1.

Table 1: Dataset Description

Feature	Description	Data Type
Age (yrs)	Age of the patient in years	Float64
Weight (Kg)	Body weight in kilograms	Float64
Height (Cm)	Height in centimetres	Float64
BMI	Body Mass Index	Float64
Blood Group	Blood type as numerical code	Int64
Pulse rate(bpm)	Pulse rate in beats per minute	Float64
RR (breaths/min)	Respiratory rate per minute	Int64
Cycle(R/I)	Menstrual cycle regularity	Int64
Pregnant(Y/N)	Pregnancy status (1=Yes, 0=No)	Int64
I beta-HCG (mIU/mL)	Beta-HCG hormone level (case I)	Float64
AMH (ng/mL)	Anti-MÃ¼llerian Hormone level	Float64
FSH (mIU/mL)	Follicle Stimulating Hormone	Float64
LH (mIU/mL)	Luteinizing Hormone	Float64
FSH/LH	Ratio of FSH to LH	Float64
Hair loss(Y/N)	Presence of hair loss (1=Yes, 0=No)	Int64
Skin darkening (Y/N)	Presence of skin pigmentation (1=Yes, 0=No)	Int64
Weight gain(Y/N)	Reported weight gain (1=Yes, 0=No)	Int64
Hair growth(Y/N)	Excessive hair growth (1=Yes, 0=No)	Int64
Pimples(Y/N)	Presence of pimples/acne (1=Yes, 0=No)	Int64
Fast food (Y/N)	Fast food consumption (1=Yes, 0=No)	Float64
Reg.Exercise(Y/N)	Engagement in regular exercise (1=Yes, 0=No)	Int64
Follicle No. (L)	Number of follicles in left ovary	Int64
Follicle No. (R)	Number of follicles in right ovary	Int64
Avg. F size (L) (mm)	Average follicle size in left ovary	Float64
Avg. F size (R) (mm)	Average follicle size in right ovary	Float64
Endometrium (mm)	Thickness of the endometrial lining	Float64
PCOS (Y/N)	Diagnosis of PCOS (1=Yes, 0=No)	Int64

3.3. Dataset Preparation

After completing the data exploration phase, the dataset undergoes a comprehensive preprocessing stage. This phase includes handling missing values, eliminating duplicate records, applying normalization, selecting relevant features, encoding categorical variables, and splitting the data into training and testing sets. These preprocessing steps are crucial to ensure the dataset is clean, well-structured, and suitable for accurate modelling and further analysis.

3.3.1. Missing Data

To ensure the integrity of the dataset, two standard validation functions were applied: isnull (). sum () and duplicated (). sum (). For instance, the isnull (). sum () function was used to detect and count missing values across all columns, while duplicated().sum() identified any repeated rows that could affect data quality. The results confirmed that the dataset contained no missing values or duplicate entries, indicating a high level of completeness and consistency. This verification step is essential, as clean and reliable data forms the foundation for developing accurate and robust ML models.

3.3.2. Balancing the Dataset

The dataset comprises a total of 541 patient records, each containing clinical, biochemical, and lifestyle-related information relevant to the diagnosis of PCOS. The target variable, PCOS (Y/N), is binary, where 1 indicates a positive PCOS diagnosis and 0 indicates the absence of the condition as presented in Figure 2. To address this imbalance and improve the performance of ML models, the study employed SMOTE. The SMOTE generates synthetic examples of the minority class (PCOS) to create a more balanced dataset. This technique helps reduce bias toward the majority class during model training, leading to more reliable and generalizable classification outcomes [17].

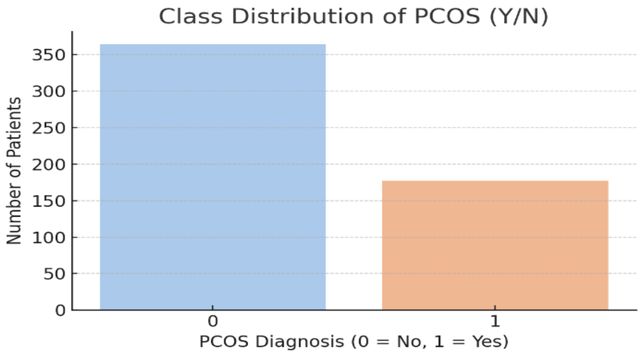

Figure 2: Class Distribution of PCOS

3.3.3. Feature Selection

The feature selection results using ANOVA F-scores highlight the most statistically significant variables for distinguishing between PCOS and non-PCOS cases. The two most predictive features are Follicle No. (R) and Follicle No. (L), with F-scores of 390.84 and 308.52, respectively. These findings are consistent with clinical criteria, as women with PCOS typically present with a higher number of ovarian follicles, particularly in the right ovary. Other highly discriminative features include skin darkening, hair growth, and weight gain, all of which are common symptoms associated with hormonal imbalance and insulin resistance in PCOS patients.

The menstrual cycle regularity feature (Cycle R/I) also shows a high F-score (103.67), emphasizing its importance, as irregular cycles are a key diagnostic marker of PCOS. Moderate contributions come from features like fast food consumption, pimples, weight, BMI, and cycle length, which reflect both lifestyle and physiological factors influencing the condition. Less predictive but still relevant features include hair loss, age, waist size, and hip circumference, which contribute to the model with lower F-scores. Overall, the analysis confirms that reproductive indicators, clinical symptoms, and lifestyle behaviours play a vital role in the classification of PCOS, guiding both feature prioritization and model development for improved diagnostic accuracy. A summary of the attribute's importance is provided in Table 2.

Table 2: Feature Importance Using ANOVA F-score

Selected Feature	ANOVA F-score
Follicle No. (R)	390.83
Follicle No. (L)	308.51
Skin darkening (Y/N)	157.67
hair growth(Y/N)	148.42
Weight gain(Y/N)	130.16
Cycle(R/I)	103.67
Fast food (Y/N)	89.72
Pimples(Y/N)	48.04
Weight (Kg)	25.34
BMI	22.34
Cycle length(days)	17.73
Hair loss(Y/N)	16.6
Age (yrs)	15.75
Waist(inch)	15
Hip(inch)	14.58

3.3.4. Encoding Categorical Data

The dataset was processed using label encoding to convert categorical variables into numerical format, a crucial preprocessing step as most ML algorithms requires numerical input [20]. In this study, all categorical features were successfully transformed into numeric values. This conversion was essential to ensure compatibility with the classification models, ultimately enhancing the efficiency and accuracy of the training and evaluation processes.

3.3.5. Splitting Data

The dataset was initially divided into two subsets, with 80% allocated for training and 20% for testing. This split enables the model to learn patterns from the larger portion of the data while using the remaining portion to assess its performance on previously unseen instances, ensuring a more reliable evaluation.

3.3.6. Data Normalization

The numerical features were normalized to scale their values within a consistent range, typically between 0 and 1. This process ensures that all features contribute equally during model training, preventing any single variable from dominating the learning process. Normalization supports more balanced and unbiased model performance, ultimately enhancing the accuracy and stability of the results [21].

3.4. Modelling

Seven ML algorithms—LR, NB, SVM, RF, GBC, AdaBoost, and MLP—were applied to classify patients based on the presence or absence of PCOS.

LR is a supervised ML algorithm commonly used for binary classification tasks. It estimates the probability that a given input belongs to a particular class by applying a sigmoid function to a linear combination of the input features. The output is a value between 0 and 1, representing the likelihood of the positive class. LR is valued for its simplicity, interpretability, and efficiency, making it a reliable choice for solving classification problems in various domains [22].

RF is an ensemble ML method that constructs numerous DTs during the training phase and combines their predictions to enhance accuracy and stability. For classification tasks, it typically uses majority voting to determine the final output. This approach helps reduce both overfitting and variance compared to relying on a single DT, leading to improved generalization and performance on new, unseen data [23].

GBC is an effective ensemble learning method that constructs models in a sequential manner, with each new model aiming to improve upon the errors of its predecessors. It combines multiple weak learners, typically shallow DTs, and optimizes performance by minimizing a loss function through gradient-based techniques. This approach often results in high predictive accuracy, although it may require more training time due to its iterative nature [23].

SVM is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that separates data points of different classes with the maximum margin. The data points closest to the hyperplane, known as support vectors, are critical in defining the decision boundary. SVM is especially effective in high-dimensional spaces and can be adapted to non-linear problems through the use of kernel functions. Its ability to handle complex relationships and avoid overfitting makes it a widely used method in ML [20].

NB is a simple, yet effective supervised classification algorithm based on Bayes' Theorem. It assumes that all features are independent of each other given the class label—an assumption known as "naive" independence. Despite this simplification, NB performs well in many real-world scenarios, particularly with large datasets. It is computationally efficient, easy to implement, and works well for both binary and multi-class classification problems, especially when the input features are categorical or conditionally independent [20].

AdaBoost is an ensemble learning algorithm that combines multiple weak classifiers, typically DTs, to form a strong classifier. It works by training models sequentially, where each new model focuses more on the errors made by the previous ones. During the training process, weights are assigned to each instance, increasing for those that are misclassified, so the next model gives them more attention. AdaBoost is known for improving accuracy, reducing bias, and being relatively resistant to overfitting when properly tuned. It performs well on binary classification tasks and is particularly effective with clean, well-prepared data [24].

MLP is a type of ANN used for supervised learning tasks, including both classification and regression. It consists of an input layer, one or more hidden layers, and an output layer, with each layer made up of interconnected nodes (neurons).

MLP uses non-linear activation functions and is trained using backpropagation to minimize prediction errors. It is capable of capturing complex patterns in the data but often requires careful tuning of hyperparameters and sufficient data to perform effectively. MLP is particularly useful when the relationship between features and outcomes is non-linear and not easily captured by simpler models [25].

3.5. *Performance Evaluation*

The performance of the supervised ML models is assessed using key evaluation metrics—accuracy, precision, recall, F-measure and ROC AUC—which together offer a comprehensive understanding of each model's classification effectiveness.

3.5.1. *Accuracy:*

It measures the proportion of correctly predicted instances out of the total number of predictions. It reflects the overall effectiveness of a model incorrectly classifying both positive and negative cases, as expressed in Equation (1) [26].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

3.5.2. *F-measure:*

It provides a balanced evaluation by combining precision and recall into a single metric. It is especially valuable when dealing with imbalanced datasets or when both false positives and false negatives carry significant consequences, as shown in Equation (2) [26].

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{2}$$

3.5.3. *Precision:*

It quantifies the ratio of correctly predicted positive instances to all instances predicted as positive. It evaluates the model's ability to produce reliable positive predictions, helping determine how many of the predicted positives are relevant. This is illustrated in Equation (3) [26].

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

3.5.4. *Recall:*

It measures the proportion of actual positive cases that are correctly identified by the model. It is crucial in contexts where missing positive cases may have serious implications, as represented in Equation (4) [26].

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

3.5.5. *ROC AUC*

It is a performance metric used to evaluate the classification ability of a ML model across various threshold settings. The ROC curve plots the True Positive Rate against the False Positive Rate, showing how the model's sensitivity and specificity vary with different decision boundaries. The AUC quantifies the overall ability of the model to distinguish between classes [26].

4. Results

The results of the current study demonstrate the effectiveness of the ML techniques in accurately predicting PCOS. Key performance metrics, including accuracy, precision, recall, F1-Score and ROC AUC, were evaluated to assess model reliability. As provided in Table 3.

Table 3: Performance Comparison Between Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
NB	90.8	82.4	87.5	84.8	96.7
LR	91.7	96.0	85.0	84.2	96.8
SVM	89.9	92.0	90.0	80.7	96.0
RF	89.0	83.3	78.1	80.6	95.0
GBC	89.0	83.3	78.1	80.6	92.1
AdaBoost	88.1	85.2	71.9	78.0	93.4
MLP	87.2	82.1	71.9	76.7	92.1

In terms of accuracy, LR achieved the highest score at 91.7%, indicating its strong overall capability to correctly classify both positive and negative cases. NB followed closely with 90.8%, while SVM and RF achieved 89.9% and 89%, respectively. GBC also matched RF with 89%, and AdaBoost recorded a slightly lower accuracy at 88.1%. The MLP had the lowest accuracy among all models at 87.2%, suggesting it may be less effective in general classification performance for this dataset as shown in Figure 3.

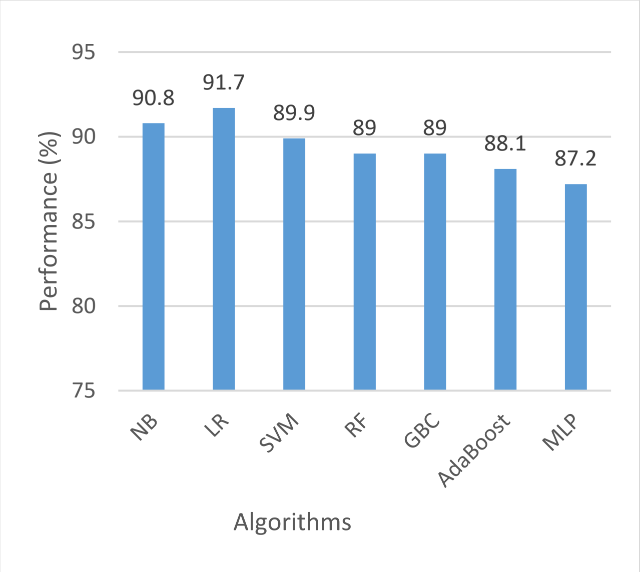

Figure.3: Accuracy Plot of Proposed Models

When evaluating precision, which measures the correctness of positive predictions, LR outperformed all other models with a precision of 96%. SVM came next with 92%, indicating its reliability in predicting relevant positive cases. AdaBoost followed with 85.2%, and both RF and GBC scored 83.3%. NB had a precision of 82.4%, and MLP was the lowest at 82.1%. This metric highlights LR as the most dependable model when minimizing false positives is important as shown in Figure 4.

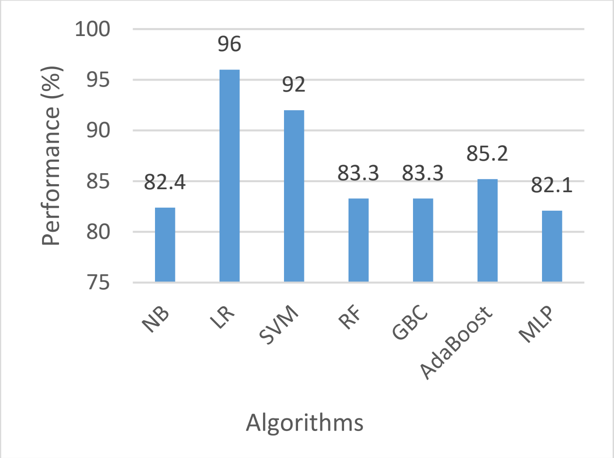

Figure.4: Precision Plot of Proposed Models

The performance comparison based on recall shows that NB achieved the highest recall at 87.5%, demonstrating superior sensitivity in correctly identifying positive cases. This is followed by LR, which also performed well with a recall of 85%, indicating its effectiveness with the dataset's linear characteristics. Meanwhile, SVM, AdaBoost, and MLP exhibited moderate recall values of 79%, reflecting balanced but less outstanding performance in detecting positive cases. Finally, RF and GBC recorded the lowest recall values at 78.1%, suggesting that these ensemble methods may have underperformed in this specific context, possibly due to data characteristics or parameter tuning limitations as shown in Figure 5.

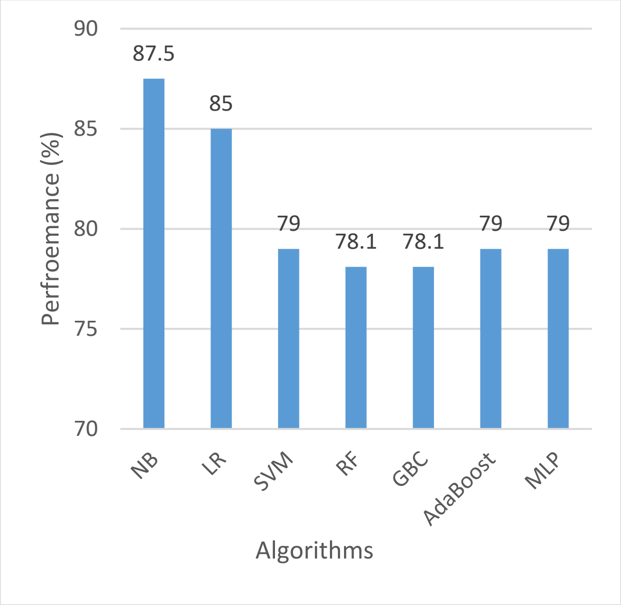

Figure.5: Recall Plot of Proposed Models

For F1 Score, which balances both precision and recall, NB again emerged as the top performer with an F1 Score of 84.8%, suggesting it offers the most balanced predictions. LR was a close second at 84.2%. SVM, RF, and GBC showed similar F1 scores around 80.6–80.7%, reflecting solid but slightly less balanced performance. AdaBoost scored 80%, while MLP had the lowest F1 Score at 76.7%, further confirming its relatively weaker balance between identifying and correctly classifying positive cases as shown in Figure 6.

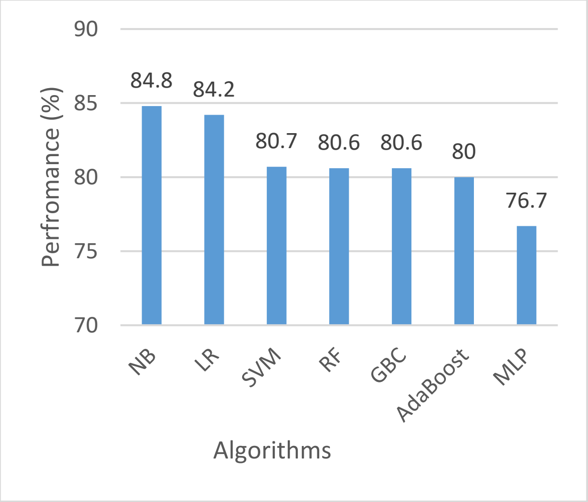

Figure.6: F1 Score Plot of Proposed Models

Regarding ROC AUC, which assesses a model's ability to distinguish between classes at various threshold levels, LR achieved the highest score of 96.8%, closely followed by NB at 96.7% and SVM at 96%. RF also performed well with 95%, and AdaBoost came next at 93.4%. The lowest AUC scores were observed in GBC and MLP, both at 92.1%. These results indicate that while all models demonstrated good class-separating ability, LR and NB were the most effective in this regard as shown in Figure 7.

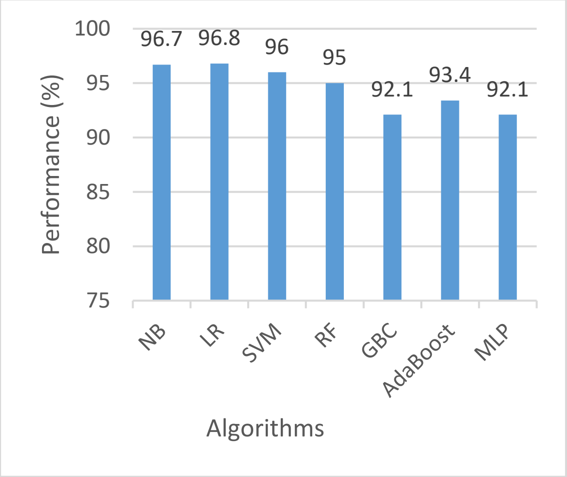

Figure.7: ROC AUC Plot of Proposed Models

5. Discussion

The superior performance of the LR model, achieving the highest AUC, aligns with findings from previous studies discussed in the literature. Similar to the work of Hosain et al. [6], where LR achieved an accuracy of 85.3% due to its strong predictive capability with hormonal and metabolic attributes, this study also demonstrated the effectiveness of LR when supported by appropriate feature selection and data balancing techniques. In the present analysis, class imbalance was effectively managed using the SMOTE algorithm, enhancing the model's sensitivity and specificity—an approach also highlighted by Shanmugavadivel et al. [15] in addressing rare class detection.

Additionally, feature selection using ANOVA F-scores helped identify the most statistically significant predictors, allowing LR to focus on the most influential clinical variables, consistent with the methodology applied by Hosain et al. [6]. These results further validate the literature's emphasis on the importance of simple, interpretable models like LR, particularly when combined with effective preprocessing strategies, achieving performance comparable to or even surpassing more complex models such as RF and SVM [5], [13].

Although the models, particularly LR, achieved high accuracy and AUC scores, we acknowledge that recall values were modest in several cases, indicating a proportion of PCOS cases were not successfully identified. This raises clinical concerns, as missed diagnoses in screening settings may delay treatment. To address this, we will conduct further analysis of false negative cases to identify potential patterns or limitations in feature representation. Additionally, we plan to experiment with threshold tuning, cost-sensitive learning, and advanced resampling methods to improve recall. In clinical contexts, high recall is essential to ensure at-risk patients are not overlooked. A comparative benchmark with clinical diagnostic rates among physicians will also be considered in future work to contextualize the model's performance

6. Conclusion a Future Direction

This study evaluated the performance of seven supervised ML algorithms— LR, NB, SVM, RF, GBC, AdaBoost, and MLP —for the classification of PCOS based on clinical and lifestyle data. The models were assessed using key performance metrics including accuracy, precision, recall, F1 score, and ROC AUC. Among all the models, LR consistently demonstrated the best overall performance.

LR achieved the highest accuracy (91.7%), precision (96%), and ROC AUC (96.8%), and maintained a strong balance between recall and F1 score. Its superior performance can be attributed to the linear separability of the dataset and the model's inherent ability to generalize well with limited assumptions and minimal overfitting. Furthermore, LR is computationally efficient, easy to interpret, and performs reliably when the relationship between features and output is approximately linear characteristics that align well with the nature of this dataset.

This study confirms the potential of machine learning (ML) in identifying PCOS with high accuracy and interpretability. However, limitations such as moderate recall scores, missing hormonal and demographic variables, and the absence of comparison with clinical decision-making indicate that the current approach requires further enhancement before clinical adoption. Addressing these gaps will improve both the diagnostic value and real-world applicability of ML models in women's health.

Future work should focus on incorporating more comprehensive clinical and biochemical indicators, including insulin resistance markers, androgen levels, and family history. Advanced ensemble techniques like XGBoost and model stacking could be employed to boost predictive performance. Additionally, combining structured data with medical imaging or exploring deep learning (DL) models may lead to more robust diagnostic tools. Expanding the dataset to include diverse populations and validating findings in clinical settings will also be key to ensuring generalizability and fairness in AI-assisted PCOS diagnosis.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The Authors hereby acknowledge that the funding of this paperwork was done and shared across all Authors concerned.

References

[1] C. C. Dennett and J. Simon, "The role of polycystic ovary syndrome in reproductive and metabolic health: overview and approaches for treatment," Diabetes Spectrum,vol. 28, no. 2, pp. 116-120, 2015. DOI: 10.2337/diaspect.28.2.116

[2] I. T. Lee et al., "Depression, anxiety, and risk of metabolic syndrome in women with polycystic ovary syndrome: a longitudinal study," The Journal of Clinical Endocrinology Metabolism,vol. 110, no.3, pp. e750-e756, 2025. DOI: 10.1210/clinem/dgae256

[3] Z. Zad et al., "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," Frontiers in Endocrinology, vol. 15, p. 1298628, 2024. DOI: 10.3389/fendo.2024.1298628

[4] C. Tong, Y. Wu, Z. Zhuang, and Y. Yu, "A diagnostic model for polycystic ovary syndrome based on machine learning," Scientific Reports,vol. 15, no. 1, p. 9821, 2025. DOI: 10.1038/s41598-025-92630-4

[5] P. Chauhan, P. Patil, N. Rane, P. Raundale, and H. Kanakia, "Comparative analysis of machine learning algorithms for prediction of pcos," in 2021 international conference on communication information and computing technology (ICCICT), 2021, pp. 1-7: IEEE. DOI: 10.1109/ICCICT50803.2021.9509757

[6] A. S. Hosain, M. H. K. Mehedi, and I. E. Kabir, "Pconet: A convolutional neural network architecture to detect polycystic ovary syndrome (pcos) from ovarian ultrasound images," in 2022 International Conference on Engineering and Emerging Technologies (ICEET), 2022, pp. 1-6: IEEE. DOI: 10.1109/ICEET56468.2022.10012345

[7] D. Rao, R. R. Dayma, and S. K. Pendekanti, "Deep learning model for diagnosing polycystic ovary syndrome using a comprehensive dataset from Kerala hospitals," International Journal of Electrical Computer Engineering.vol. 14, no. 5, 2024. DOI: 10.11591/ijece.v14i5.36503

[8] B. Panjwani, J. Yadav, V. Mohan, N. Agarwal, and S. Agarwal, "Optimized Machine Learning for the Early Detection of Polycystic Ovary Syndrome in Women," Sensors.vol. 25, no. 4, p. 1166, 2025. DOI: 10.3390/s25041166

[9] L. Ji et al., "Performance of a Full-Coverage Cervical Cancer Screening Program Using on an Artificial Intelligence–and Cloud-Based Diagnostic System: Observational Study of an Ultralarge Population," Journal of Medical Internet Research.vol. 26, p. e51477, 2024. DOI: 10.2196/51477

[10] M. Arya, "Automated detection of acute leukemia using K-means clustering algorithm," 2019. DOI: 10.5120/ijca2019918801

[11] H. O. Boll et al., "Graph neural networks for clinical risk prediction based on electronic health records: A survey," J. Biomed. Informatics.vol. 151, p. 104616, 2024. DOI: 10.1016/j.jbi.2024.104616

[12] M. de Oliveira Gomes, J. de Oliveira Gomes, L. F. Ananias, L. A. Lombardi, F. S. da Silva, and A. P. Espindula, "ANTI-MÜLLERIAN HORMONE AS A DIAGNOSTIC MARKER OF POLYCYSTIC OVARY SYNDROME: A SYSTEMATIC REVIEW WITH META-ANALYSIS," American Journal of Obstetrics Gynecology,2025. DOI: 10.1016/j.ajog.2025.03.077

[13] M. Wang et al., "Biochemical classification diagnosis of polycystic ovary syndrome based on serum steroid hormones," The Journal of Steroid Biochemistry Molecular Biology,vol. 245, p. 106626, 2025. DOI: 10.1016/j.jsbmb.2024.106626

[14] K. M. Mohi Uddin, M. T. A. Bhuiyan, M. M. Rahman, M. M. Islam, and M. A. Uddin, "Early PCOS Detection: A Comparative Analysis of Traditional and Ensemble Machine Learning Models With Advanced Feature Selection," Engineering Reports,vol. 7, no. 2, p. e70008, 2025. DOI: 10.1002/eng2.70008

[15] K. Shanmugavadivel, M. D. MS, M. TR, T. Al-Shehari, N. A. Alsadhan, and T. E. Yimer, "Optimized polycystic ovarian disease prognosis and classification using AI based computational approaches on multi-modality data," BMC Medical Informatics Decision Making, vol. 24, no. 1, p. 281, 2024. DOI: 10.1186/s12911-024-02688-9

[16] T. Zohrabi, A. Nadjarzadeh, S. Jambarsang, M. H. Sheikhha, A. Aflatoonian, and H. Mozaffari-Khosravi, "Effect of dietary approaches to stop hypertension and curcumin co-administration on glycemic parameters in polycystic ovary syndrome: An RCT," International Journal of Reproductive BioMedicine,vol. 22, no. 9, p. 689, 2024. DOI: 10.18502/ijrm.v22i9.14994

[17] R. Ahmad, L. A. Maghrabi, I. A. Khaja, L. A. Maghrabi, and M. Ahmad, "SMOTE-Based Automated PCOS Prediction Using Lightweight Deep Learning Models," Diagnostics, vol. 14, no. 19, p. 2225, 2024. DOI: 10.3390/diagnostics14192225

[18] F. Ahmad Musleh, "A Comprehensive Comparative Study of Machine Learning Algorithms for Water Potability Classification," International Journal of Computing Digital Systems,vol. 15, no. 1, pp. 1189-1200, 2024. DOI: 10.12785/ijcds/150112

[19] P. Kottarathil, "Polycystic Ovary Syndrome (PCOS)," https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos, May 10, 2025 2020. DOI: 10.34740/KAGGLE/DSV/1203444

[20] S. Alshakrani, R. Taha, and N. Hewahi, "Chronic kidney disease classification using machine learning classifiers," in 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2021, pp. 516-519: IEEE. DOI: 10.1109/3ICT51146.2021.9589834

[21] F. A. Musleh and R. G. Taha, "Forecasting of forest fires using machine learning techniques: a comparative study," 2022.

[22] R. Taha, S. Alshakrani, and N. Hewahi, "Exploring Machine Learning Classifiers for Medical Datasets," in 2021 International Conference on Data Analytics for Business and Industry (ICDABI), 2021, pp. 255-259: IEEE. DOI: 10.1109/ICDABI53623.2021.9655862

[23] F. Musleh, R. Taha, and A. R. Musleh, "Comparative Analysis of Machine Learning Techniques for Concrete Compressive Strength Prediction," in 2023 4th International Conference on Data Analytics for Business and Industry (ICDABI), 2023, pp. 146-151: IEEE. DOI: 10.1109/ICDABI60145.2023.10629479

[24] E. Dritsas and M. Trigka, "Efficient Data-Driven Machine Learning Models for Water Quality Prediction," Computation,vol. 11, no. 2, p. 16, 2023. DOI: 10.3390/computation11020016

[25] X. Hu, A. Yadav, A. Khan, A. P. Sah, and S. Azam, "Construction of PCOS Prediction Model Based on BP Neural Network," in 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 2025, pp. 885-889: IEEE. DOI: 10.1109/ICMCSI64620.2025.10883524

[26] A. M. Zeki, R. Taha, and S. Alshakrani, "Developing a predictive model for diabetes using data mining techniques," in 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2021, pp. 24-28: IEEE. DOI: 10.1109/3ICT53449.2021.9582114

Mrs. Ranyah Taha completed her MSc in Big Data Science and Analytics in 2022 through a joint program between Liverpool John Moores University and the University of Bahrain. She earned her BSc in Computer Science from the University of Bahrain in 2018. Her research focuses on leveraging Data Science and Analytics, particularly Machine Learning and Deep Learning, to build advanced models and extract valuable insights from complex datasets. She has contributed to many research papers and was awarded the NASA International Space Apps Challenge – Space Apps Bahrain 2023 Local Impact Award.

Miss Huda Zain El Abdin completed her MSc in Data Science with distinction in 2025. Her graduation project was recognised as one of the top five best graduation projects of the year. She earned her BSc in Software Engineering from the University of Bahrain in 2021. Her research interests lie in the development and application of advanced Natural Language Processing techniques, including LLMs, to solve real-world language understanding challenges. She is particularly interested in the intersection of machine learning and the medical field, exploring how AI can enhance healthcare delivery and diagnostics. She was also awarded second place in a NLP Hackathon organised by London Business School and Middlesex University.

Miss Tala Musleh Pharmacy student at the University of Bahrain, with a keen interest in applying Artificial Intelligence (AI) and Machine Learning (ML) to advance research and clinical practices in the medical and pharmaceutical field.