I.J. Information Technology and Computer Science, **2025, 1, 82-95**
Published Online on February 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2025.01.06

NIPP: Non-Invasive PCOS Prediction using XG-boost Machine Learning Model

Shikha Prasher
Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India
E-mail: shikhamalhan1803@gmail.com
ORCID iD: https://orcid.org/0000-0002-8732-0239

Leema Nelson*
Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India
E-mail: leema.nelson@gmail.com
ORCID iD: https://orcid.org/0000-0003-3959-5390
*Corresponding Author

Manal Gafar
Cyber Security and Networks Program, University of East London, European Universities in Egypt (EUE), The New Administrative Capital, Egypt
Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt
E-mail: Manal.amr@eue.edu.eg
ORCID iD: https://orcid.org/0009-0008-6717-7552

Received: 27 August 2024; Revised: 13 October 2024; Accepted: 02 December 2024; Published: 08 February 2025

Abstract: Polycystic Ovary Syndrome (PCOS) is a common endocrine disorder that affects women of reproductive age, leading to hormonal imbalances and ovarian dysfunction. Early detection and intervention are vital for effective management and prevention of complications. This study compares PCOS prediction using the XGBoost machine learning model against four traditional models: Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees (DT), and Random Forests (RF). LR and SVM achieve accuracies of 95% and 96%, respectively, demonstrating strong predictive capabilities. In contrast, DT had a lower accuracy (82%), indicating limitations in PCOS data complexity. RF showed competitive performance with 96% accuracy, underscoring its effectiveness in ensemble learning. XGBoost achieves 98% accuracy with its parameter configuration. The scale pos weight parameter adjusts the positive class weight in imbalanced datasets, addressing under representation by assigning more weight to the minority class, and thereby improving the training focus. The gradient boosting framework incrementally builds models to address complex feature interactions and dependencies, enhancing the accuracy and stability in predicting intricate PCOS dataset. This analysis highlights the importance of advanced machine learning models such as XGBoost for accurate and reliable PCOS predictions. This research advances PCOS prediction, demonstrates the potential of machine learning in healthcare, and clarifies the strengths and limitations of different algorithms with complex medical datasets.

Index Terms: Machine Learning, Polycystic Ovary Syndrome, Extreme Gradient Boosting, Numeric Database, Early Detection.

1. Introduction

PCOS is known as polycystic ovary syndrome or" Syndrome O" (overnutrition, over insulin production, ovarian dis-orientation, and ovulation). Symptoms of polycystic ovaries, irregular or non-existent menstrual periods, and elevated androgenic hormones might manifest as PCOS physiologically or hormonally [1]. PCOS is a medical illness characterized by enlargement of the ovaries in females, accompanied by a notable prevalence of cysts. It is important to note that PCOS is classified as a syndrome rather than a disease. The cysts observed are characterized by immature follicles. The occurrence of anovulation, which refers to the failure of mature follicles to be released, is increasingly inhibited as the disorder progresses because of thickening of the menstrual wall. PCOS is characterized by several

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

clinical manifestations in women with polycystic ovaries, including obesity, hirsutism, menstrual cycle irregularities, acne, and abnormal biochemical features. These features often lead to elevated testosterone, insulin, androstenedione, and blood luteinizing hormone levels. In addition, women with PCOS may experience unanticipated weight gain and accelerated hair growth. Insulin resistance is a notable determinant that influences the pathophysiology of PCOS [2]. According to doctors, ovarian cysts do not rank among the primary etiological factors or diagnostic criteria for PCOS. Based on prior scholarly investigations, it has been observed that a significant proportion (30–70 percentages) of females diagnosed with PCOS manifest obesity as their principal sign. This finding implies a reciprocal association between PCOS and obesity. In addition to cystic ovaries and abnormal periods, increased testosterone synthesis is recognized as a contributing component in the manifestation of PCOS [3]. Infertility, diabetes mellitus, reproductive dysfunction, hyperinsulinemia, obesity, and breast cancer are significant contributors to global mortality. PCOS is a prominent etiological factor contributing to infertility, affecting a substantial proportion of women in India, with prevalence rates reaching up to 20% [4].

1.1. Symptoms of PCOS

People with ovaries of reproductive age are the majority of those affected by PCOS, a common metabolic condition. Multiple features and symptoms of PCOS, such as irregular menstruation cycle, ovarian disorder, hyperandrogenism, insulin resistance, obesity, weight gain, skin issues, and mood changes, may appear, and not everyone with PCOS will have a similar set of signs. An irregular menstruation cycle refers to menstrual periods that are irregular or infrequent and are frequently caused by PCOS. Some patients with PCOS may experience fewer than eight menstrual cycles annually. Many people with PCOS have difficulty ovulating (emitting eggs from the ovaries) consistently, which can lead to fertility issues, which are known as ovarian disorders. Male hormones called androgens, such as testosterone, can induce symptoms including acne, excessive body or facial hair, and male-pattern baldness [5]. PCOS patients may exhibit larger ovaries on ultrasound along with numerous small fluid-filled sacs called follicles. Cysts are immature eggs that are not ovulated or expelled. Insulin levels in the blood may increase because many PCOS patients have insulin resistance. This could increase weight gain and the risk of type 2 diabetes, making it more difficult to maintain healthy weight. An elevated risk of weight gain and obesity is associated with PCOS. Having too much weight might worsen the signs of PCOS and increase the possibility of complications. Individuals with PCOS may experience skin issues such as acne, oily skin cells, and tags on the skin more frequently [6]. Anxiety and depression are among the mental health issues that some people with PCOS may suffer.

1.2. Need for PCOS Prediction

The identification of PCOS is important for various reasons, and the early detection of PCOS enables prompt intervention and treatment. Timely detection can aid in the effective management of symptoms and mitigate the likelihood of developing cardiovascular ailments, type 2 diabetes and infertility [7]. PCOS symptoms can occur in many different forms, including irregular menstrual cycles, hirsutism, acne, and weight gain. The precise identification of medical conditions enables healthcare practitioners to customize treatment strategies to efficiently address these symptoms. Ovulation irregularities are a major cause of infertility in many women with PCOS patients. Early detection can help determine the best reproductive treatment and options that improve the likelihood of a successful birth [8]. PCOS frequently correlates with insulin resistance and elevated susceptibility to the onset of type 2 diabetes. Timely identification of insulin resistance and implementation of lifestyle adjustments can effectively mitigate its impact and potentially forestall or postpone the development of diabetes [9]. The early identification of PCOS enables healthcare providers to impart knowledge regarding potential health hazards to individuals and devise effective ways to mitigate these risks. This encompasses the adoption of a health-conscious lifestyle, effective weight management, and the proactive treatment of other interconnected health issues. The precise cause of PCOS is indeterminate; however, it is believed to be the integration of genetic, hormonal, and external variables. PCOS is associated with a higher risk of diabetes, obesity, heart disease, and infertility. The condition is persistent and requires continuous treatment [10]. PCOS is typically diagnosed through a combination of clinical signs, physical checkups, blood tests to evaluate hormone levels, and ultrasound scans of the ovaries. Normal PCOS treatment strategies include dietary and exercise improvements, hormonal remedies to modulate menstrual cycles and regulate symptoms, and the management of associated medical conditions.

1.3. Non-invasive PCOS Prediction

Artificial Intelligence (AI) encompasses various cognitive abilities, such as thinking, learning, problem-solving skills awareness, and language understanding. AI has sparked a significant revolution and holds immense potential for the future. It encompasses the development of robots that emulate human cognitive processes, including learning and problem- solving. The proliferation of artificial intelligence (AI) applications is widespread in various domains. AI has the potential to efficiently and accurately address real-world challenges. AI is increasingly employed by numerous healthcare businesses to enhance outcomes and reduce instances of human mistakes. Numerous researchers have employed AI techniques to categorize ultrasound images autonomously. AI can acquire knowledge from extensive datasets in healthcare settings, enabling it to identify diseases effectively. AI can generate outcomes with reduced or negligible errors and eliminate or discard extraneous and nonessential data from its operations [1]. PCOS can be detected using AI techniques such as Machine Learning (ML) and Deep learning (DL) models.

2. Literature Review

CICEK et al. [11] have proposed that PCOS patient' risk factors are identified using the Local Interpretable Model-agnostic Explanations (LIME) method, which uses an explainable AI algorithm in 2021. This research focuses on the extraction of significant features for decision-making using the local interpretable model-agnostic annotation (LIME) approach. PCOS risk variables were understood and estimated using the RF model. The balanced accuracy obtained by RF using the LIME method was 85.84%. Roy et al. [12] have proposed Automation in Polycystic Ovarian Syndrome Diagnosis in 2021. An SVM classifier was employed in this study to diagnose patients with PCOS. The sample included 541 females under the age of 47 years who had both positive and negative class designations. The SVM outperformed, with a prediction accuracy of 96.22%. Rathod et al. [13] have proposed utilising the CatBoost algorithm for predictive analysis of PCOS in 2022.In this study, ML techniques including KNN, SVM, RF, NB, CatBoost, and LR were thoroughly examined and utilized for classification problems. CatBoost stands out among them because of its exceptional performance; in some cases, it can achieve up to 96% accuracy. Marreiros et al. [14] have proposed machine learning- based classification of PCOS using correlation weight in 2022. In this study, ML algorithms demonstrated the efficacy of Artificial Neural Networks (ANN), RF, DT, NB, and LR, for categorization problems. RF was one of these algorithms that has shown to be a reliable and accurate technique; in certain experiments, it has been shown to have an accuracy rate of 93.06%. Mehr et al. [15] have proposed diagnosis of PCOS using various machine learning and feature selection methods in 2022. In this study, the effectiveness and performance of machine learning algorithms for classification tasks, including Random Forest (RF), Extra Trees, AdaBoost, and MLP, were studied. With an astounding accuracy rate of 98% in certain experiments, Random Forest stands out among these algorithms, demonstrating its resilience and capacity to manage complicated datasets well.

Bhardwaj et al. [16] proposed the utilization of machine learning algorithms in the healthcare sector for PCOS diagnosis in 2022. In this study, RF, SVM, XGBoost, and MLP are machine-learning methods for classification tasks that have been thoroughly examined and contrasted. SVM outperformed the other methods with a 93% accuracy rate, proving its usefulness in classifying data using the best hyperplanes in high-dimensional spaces. SVM is a popular option in many applications, such as image recognition and bioinformatics, owing to its capacity to handle nonlinear decision boundaries through kernel approaches. MLP, on the other hand, demonstrated its ability to use neural network topologies to identify intricate patterns and correlations inside data, achieving an accuracy of 91%. XGBoost, which is well known for its ensemble learning methodology, has demonstrated competitive performance using gradient boosting methods to increase prediction accuracy. Sinthia et al. [17] have proposed PCOS analysis and a comparative analysis of different machine learning algorithms in 2022. In this study, SVM, KNN, K-means clustering, and LR are machine learning methods for classification problems, whose performance has been analyzed and compared. With an accuracy of 91%, the SVM proved to be a notable performer by taking advantage of its capacity to identify the best hyperplanes for class separation in high- dimensional environments. Despite being intuitive, KNN showed a reduced accuracy of approximately 75%, possibly because of dimensionality problems and noise sensitivity. With an accuracy of 72%, K-means clustering, which was first created for unsupervised learning tasks, showed limits when used in classification settings. With an accuracy of 85%, the linear logistic regression model demonstrated its ease of use and interpretability. However, its use may be restricted because of the presumption of linear correlations between target variables and characteristics.

Subha et al. [18] have proposed PCOS computerized diagnosis employing swarm intelligence and machine learning methods in 2022.In this study, RF classifiers have been assessed and contrasted with a variety of machine learning techniques such as the chi-square test, Recursive Feature Elimination (RFE), correlation ranking, Filter Feature (FF), and particle swarm optimization (PSO). The outcomes demonstrate how well these feature selection strategies work to increase classification accuracy. In particular, the PSO accuracy of 93.87% was the greatest, highlighting its capacity to optimize feature subsets for RF models. With an accuracy of 92.64%, the chi-square test and correlation ranking also performed well, demonstrating their usefulness in finding pertinent features based on statistical dependencies and correlations with the target variable. With a respectable accuracy of 91.41%, the RFE demonstrated its capacity to iteratively choose the most useful characteristics. The results of the FF technique were competitive at 92.64%.

Rakshitha et al. [19] proposed the use of a hybrid machine learning approach, which classifies patients for the purpose of predicting PCOS by 2022. Using the advantages of both SVM for nonlinear classification and LR for probabilistic interpretation and feature importance analysis, the hybrid SVM with the LR technique showed a promising accuracy of 89.03%. Tiwai et al. [20] have proposed in 2022 In this study, Non-invasive screening criteria were used in conjunction with a range of machine learning algorithms to screen PCOS patients. The out-of-bag error was employed to assess the precision of the RF forecasts, the out-of-bag (OOB) error is employed. RF approach outperformed the existing ML algorithms with an accuracy of 93.25%. Templeman et al. [21] have proposed acoustic process monitoring measurements for the identification of keyhole pore forms in laser powder-bed fusion in 2022In this Study, in which SVM and other machine learning approaches for signal processing and classification have been used in conjunction with more complex signal decomposition techniques such as Fast Fourier Transform (FFT) and Ensemble Empirical

Mode Decomposition (EEMD). With an astounding 97% accuracy rate, SVM proved its usefulness in signal classification from EEMD and FFT studies. SVM can accurately categorize complicated patterns when nonstationary and nonlinear signals are broken down into simpler components, which is made possible by EEMD. However, FFT plays a useful role in converting data into frequency domains, providing SVM frequency-based features for classification.

Khanna et al. [22] proposed a unique explainable architecture for machine learning to diagnose polycystic ovarian syndrome in 2023.In this study, a range of well-known machine-learning algorithms and feature-selection techniques were assessed for classification tasks, including DT, RF, LR, XGBoost, KNN, SVM, and ADABoost. Among these, the Salp Swarm Algorithm (SSA) feature selection technique performed exceptionally well, with a maximum accuracy of 98%. SSA efficiently finds and chooses the most informative features for classification, thereby improving the accuracy and efficiency of the model. This was inspired by the swarming behavior of the salps. Suha et al. [23] have proposed Investigated the predominant characteristics and employing a modified stacking ensemble machine learning technique to detect polycystic ovarian syndrome based on data in 2023.Multiple boosting algorithms such as AdaBoost, XGBoost, RF, Categorical Boosting, and Gradient Boosting have been investigated and compared in ensemble learning techniques for classification tasks. By gradually training weak learners to reduce prediction mistakes, Gradient Boosting is one of the algorithms that stands out the most, achieving an outstanding accuracy rate of 95.7%. By iteratively optimizing the model’s performance, gradient boosting uses gradient descent optimization, a technique well-known for its ability to manage complicated datasets and nonlinear interactions.

3. Materials and Methods

This section discusses the general approach and methods such as data collection, data preprocessing, data splitting, different machine learning models training and performance evaluation used to perform PCOS prediction in detail.

3.1. Dataset Description

Raw PCOS data are acquired from the Kaggle Repository, which is collected from a variety of sources including medical records, surveys, and research projects [24]. The numerical dataset consists of 541 samples, of 541 samples 177 are affected by PCOS and 364 are not affected by PCOS. All physiological and medical factors used to determine PCOS and related problems with infertility are represented by 42 input features, and one output feature is a class label.

3.2. Different ML Models used for PCOS Prediction

Several machine-learning models have been employed to determine the presence of PCOS. In this study, the XGBoost model is used to predict PCOS, and its performance is compared with that of other models including Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Decision Tree (DT). The XGBoost model demonstrated superior accuracy in predicting PCOS compared with the aforementioned models.

A. Logistic Regression (LR)

Logistic regression is a simple yet effective model architecture that is mostly applied to binary classification applications. Logistic regression is essentially a single layer with weights assigned to the input features. Each input characteristic is multiplied by the corresponding weight, and a bias term is added to the weighted values. The linear combination of the input features is represented by this weighted sum, which captures their contributions to the prediction. However, the activation function of logistic regression, typically a sigmoid or logistic function, distinguishes it. The linear combination of the input features is transformed by this activation function into a probability value between 0 and 1, which indicates the possibility of falling into a positive class.

B. Support Vector Machine (SVM)

A powerful supervised learning method for situations involving both regression and classification is the Support Vec- tor Machine (SVM). In a high-dimensional feature space, the idea behind the SVM architecture is to identify an ideal hyperplane to separate the data points of various classes. Fundamentally, support vector machines (SVMs) seek to locate the hyperplane with the largest margin or the greatest distance between the closest data points belonging to various classes. To define the decision boundary, support vectors, which are the nearest data points, are essential. SVM functions as a feature vector, where each data point is represented. The architecture of the system is as follows. The SVM algorithm minimizes the margin between classes and penalizes misclassifications to identify the hyperplane through the iterative optimization of a cost function. An input feature combination that is linear and weighted by coefficients established during training defines the decision boundary, which is also known as the hyperplane. The use of kernel functions is a funda- mental component of SVM because it enables the implicit mapping of input data into a higher-dimensional space where a linear decision boundary can be located. SVM is highly flexible for a variety of classification applications because of this method, called the kernel trick, which allows it to successfully handle nonlinear decision boundaries.

C. Decision Tree (DT)

Decision trees are versatile and accessible supervised learning methods that can be employed in both classification and regression tasks. They offer a high degree of adaptability and ease of use, making them popular choices for a wide range of applications. Whether classifying data or performing regression analysis, decision trees can provide accurate results in a straightforward and intuitive manner. The nodes in their architecture represent the features, branches for decisions, and leaf nodes for the results. Their architecture is hierarchical. The dataset is divided into distinct classes or groups by the root node at the top of the tree, which uses the characteristics that best divide the data. The final projected outcome or class label is represented by the leaf nodes; however, each internal node indicates a choice based on a particular trait that causes additional splits in the data.

D. Random Forest (RF)

The goal of Random Forest is to enhance the performance and generalization of individual decision trees through the use of ensemble learning techniques based on the decision tree architecture. Multiple decision trees that have been trained on random selections of both data and features constitute the Random Forest architecture. Based on a majority vote (for classification) or an average (for regression) of all individual decision tree predictions, each Random Forest decision tree acquires training independently and produces predictions. Every decision tree in the Random Forest separately generates a forecast for the input data point throughout the prediction process. The final forecast for the classification tasks is the average of all individual tree predictions. For regression tasks, the final prediction is determined by a majority vote among all decision trees. Compared to individual decision trees, this ensemble method helps lower the variance and enhances the generalization performance.

E. XGBoost

Extreme Gradient Boosting (XGBoost) is a powerful ensemble learning method, specifically, a gradient boosting machine. It sequentially trains weak learners, such as decision trees, each correcting their predecessor's errors. This iterative process refines predictions based on high-error areas and adapts them to complex datasets [25]. For imbalanced datasets, such as those in PCOS prediction, the XGBoost's scale pos weight parameter adjusts the model's sensitivity to the minority class. Proper weighting enhances the model's recall and sensitivity and better identifies difficult-to-classify cases [26]. XGBoost's decision trees can handle complex relationships by intricately partitioning the feature space and capturing nonlinear interactions. This capability suits medical datasets, where features interact in nonlinear way [27]. By focusing on the errors of previous models, XGBoost identifies subtle data patterns and improves the accuracy. To prevent overfitting, XGBoost uses L1 and L2 regularization, penalizes overly complex models, and focuses on truly predictive features [28]. Additional techniques, such as column sampling and shrinkage pruning, further stabilize the model, especially in high-dimensional datasets prone to overfitting. XGBoost's gradient boosting framework sequentially combines models in an ensemble using a weighted sum of predictions to minimize overall loss [29]. Each new learner targets the remaining errors, capturing the complex feature interactions that simpler models miss. This method enhances accuracy, scalability, and computational efficiency, making XGBoost ideal for predicting complex conditions, such as PCOS.

4. System Framework

The PCOS prediction system maintains a systematic flow procedure, beginning with the collection of raw PCOS data and ending with the prediction of PCOS diagnosis using machine learning models. Beyond this system flow process, the PCOS prediction system successfully preprocesses data, trains machine-learning models, evaluates their performance, and provides accurate PCOS diagnoses, as shown in figure1.

4.1. Data Preprocessing

The PCOS dataset contained no missing values; however, it was apparent that certain values constituted outliers in some instances. For example, the pulse rate of a patient was 13 beats per minute, diastolic pressure was 8, and systolic pressure was 13. The presence of outliers in a dataset can potentially distort the performance of a classifier. In this study, outlier values are identified by utilizing the z-score of each observation of the continuous-valued features. The z-score value of an observation x is calculated using equation (1).

$$Z_{score}(x) = \frac{x-\mu}{\sigma} \tag{1}$$

where μ is the mean value of the observation and σ is the standard deviation of the observation from the arithmetic mean. The established criterion states that" Any z-score greater than 3 or less than 3 is considered an outlier' [30]. The instances with an absolute z-score greater than 3 for any feature were removed from the dataset. The normalization technique, such as min-max scaling, is used to scale the characteristics to a range between 0 and 1. This is achieved by applying the min-max scaling equation to each feature x, which converts the initial feature values into a new scaled range.

$$x_{scale}(x) = \frac{x - min(x)}{max(x) - min(x)} \tag{2}$$

where x is the single-feature value. The lowest value of the feature throughout the dataset is denoted by min(X). The maximum value of the features for the entire dataset is represented by max(X). Using this transformation, the scaled feature is guaranteed to fall between 0 and 1.

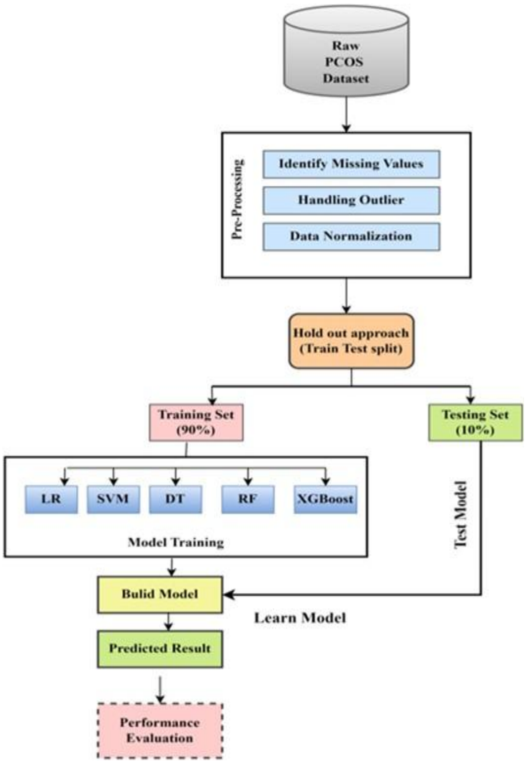

Fig.1. Work flow diagram of machine learning models

4.2. Data Splitting and Model Training

The preprocessed dataset is divided into training and testing sets using the hold-out method. Ninety percent of the data are used to train the machine learning models, and 10% of the data are used to evaluate the model’s efficiency on previously unseen data. For training, different machine-learning models, including Logistic Regression, SVM, Decision Tree, Random Forest, and XGBoost, are used. Each model is trained with training data, which included the features and labels required to learn the relationships and patterns of the data. The hyperparameters used to train the different machine learning models used in PCOS prediction are shown in Table 1.

Table 1. Hyperparameter values for models used in PCOS prediction

Model	Hyperparameter	Best Value
XGBoost	Learning Rate Maximum Tree Depth, Number of Estimators, Subsample Ratio L1 Regularization (alpha) L2 Regularization (lambda)	0.1 6 100 0.8 0.1 1
Random Forest	Number of Trees Maximum Depth, Minimum Samples per Leaf Criterion for Splitting	100 10 2 ’gini’
SVM	Choice of Kernel Regularization Parameter (C), Kernel-Specific Parameters (Gamma)	’rbf’ 1 0.1
Logistic Regression	Regularization Strength (C)	1
Decision Tree	Maximum Depth Minimum Samples Split	5 2

4.3. Model Evaluation

After training, the learned models are evaluated based on previously unseen data, using the testing set. The performance assessment criteria for each model included the accuracy, loss, precision, and recall. After the models are trained and tested, the best performing model is selected based on the evaluation metric. The selected model is then applied to predict results on fresh or previously unknown data, resulting in PCOS diagnosis estimates based on feature inputs. The system's final output includes estimated PCOS diagnoses for new instances as well as performance evaluation metrics that evaluate the system's reliability and accuracy.

5. Results and Discussion

In the Results section, the comparison analysis of PCOS prediction accuracy achieves by XGBoost and four existing machine learning models, which highlights XGBoost's better performance and investigating the potential of advanced machine learning approaches for revolutionizing PCOS detection and directing future research in this area.

5.1. Performance Evaluation

Performance metrics, including accuracy, loss, precision, and recall, are used to assess the performance of machine learning models, including those used to predict PCOS. Accuracy quantifies the extent to which forecasts align with actual outcomes.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative} \tag{3}$$

$$Presision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5}$$

$$F1\_Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

5.2. Clinical Relevance of Performance Metrics in PCOS Diagnosis

For a robust evaluation, it is essential to thoroughly examine the rationale for selecting these metrics and their relevance to the clinical context of PCOS diagnosis. Accuracy alone can be misleading, especially with imbalanced datasets where non-PCOS cases outnumber PCOS cases. This may not adequately distinguish between false positives and false negatives, which have distinct clinical implications. Moreover, Precision is crucial for PCOS diagnosis, as it measures the proportion of correctly identified PCOS cases among those diagnosed, helping to avoid overdiagnosis and its associated risks, such as unnecessary medical interventions and psychological distress. Conversely, recall reflects the model's ability to correctly identify all PCOS cases, ensuring timely treatment, and preventing exacerbation of the condition. This paper emphasizes recall because of the significant impact of missed diagnoses on a patient's quality of life and future health outcomes. The F1-score, which balances precision and recall, is valuable when there is a trade-off between these metrics and when the class distribution is uneven. It provides a single measure that reflects both the avoidance of false alarms and the assurance that no PCOS cases are missed, which is critical in healthcare. Moreover, it is important to evaluate the generalizability of the model across diverse patient populations by using independent datasets from various clinical settings. Cross-population testing based on demographic factors, such as age, ethnicity, and geographical location, is essential to ensure robustness in heterogeneous real-world scenarios. Future research will integrate the model into clinical workflows for real-time validation during patient diagnoses, incorporating clinician feedback to refine performance and align with clinical decision-making processes. Subsequent studies will compare the model's performance with conventional methods, such as ultrasound or hormone testing, to assess its added value. A hybrid approach is explored, in which machine learning identifies high-risk cases for further examination. Beyond accuracy, validation will assess the model's impact on clinical outcomes such as diagnosis time, test orders, and patient results. The cost-effectiveness analysis determines whether the model can reduce unnecessary tests and optimize workflows. Adherence to healthcare regulations, certification pathways, patient safety protocols, and informed consent procedures will be prioritized in future studies.

Acknowledging that healthcare data evolve, establish mechanisms for continuous performance monitoring and update post-deployment to maintain accuracy and relevance.

The ratio, referred to as equation (2), represents the proportion of true positive and true negative outcomes relative to the total number of conceivable cases. Loss, also known as a loss or cost function, is a quantitative measure that evaluates the disparity or error between the anticipated output of a model and actual values of the targets. However, the loss function should be lower in machine learning. A corresponding reduction in the loss values indicates that the predictions made by the model are in closer alignment with the true values, indicating an improved performance [31]. There are two methods for predicting loss: MAE and MSE. MAE calculates the average absolute difference between the expected and actual values. The MSE is a statistical metric that quantifies the average squared discrepancies between the anticipated and observed values. The algorithm assigns greater importance to larger errors, and may be influenced by extreme values. Precision is a statistic used to assess the effectiveness of a classification model, particularly for binary classification problems. It calculates the fraction of accurately predicted positive cases (true positives) among all cases projected as positive independent of the actual class. Precision is related to the model's ability to make correct positive predictions. A high precision indicates that the model makes fewer incorrect positive predictions, which leads to a lower rate of misclassifying negative cases as positive. In mathematics, the precision can be determined using equation (3). Recall, also known as the sensitivity or true positive rate, is a statistic used to assess the effectiveness of a classification model, particularly in binary classification problems. It calculates the percentage of accurately predicted positive cases (true positives) among all the actual positive cases in the dataset. Recall can be determined using Equation (4). The F1- score is a widely used metric for assessing the performance of classification models, particularly in binary classification problems. The F1-score ranges from 0 to 1, with higher values indicating greater model performance. A perfect F1- score of 1 indicated excellent accuracy and recall, whereas a score of 0 indicated poor performance. The F1-score is a comprehensive measure of a model's accuracy that considers both precision and recall. It is particularly useful for evaluating classification models with class imbalances or disproportionate misclassification rates.

True Positives are times when the outcome is correctly predicted as positive, False Positives are occasions in which something is incorrectly forecasted as positive when it is actually negative, False Negatives are occasions in which some- thing is incorrectly forecasted as negative when it is actually positive, and True Negatives are instances that have been correctly forecasted as negative. The logistic regression model for predicting PCOS showed promising results, with an overall accuracy of 95%. The model correctly identified all Class 0 cases, leading to a high F1-score of 96%. For Class 1 cases, the model demonstrated 100% precision, indicating that all predicted positive cases are indeed patients with PCOS. However, the recall is 83%, suggesting room for improvement in identifying patients with PCOS. The SVM model per- formed exceptionally well in identifying PCOS cases, with 96% accuracy rate, 100% precision, and 89% recall. The Decision Tree model exhibited reduced performance, with lower precision, recall, and F1-score values of 75%, 67%, and 71%, respectively. The model achieved a 67% recall rate, identifying 75% of positive cases as PCOS patients, but missing some true cases. With an F1-score of 71%, the overall performance of identifying positive cases is moderate. The model had an 82% accuracy rate, but poor performance in identifying PCOS patients. In contrast, the Random Forest model for PCOS prediction displayed outstanding performance in both groups, achieving 97% precision, recall, and F1-score values for individuals without PCOS (Class 0) and 94% precision, recall, and F1-score values for those with PCOS (Class 1). The model correctly identified 94% of the predicted positive cases as patients with PCOS and 94% of the actual PCOS cases. The Random Forest model had good precision, recall, and F1-score values for both classes, with a confusion matrix showing TP, TN, FP, and FN values of 37, 17, 1, and 1, respectively. The XGBoost model also performed exceptionally well in both classes, achieving 98% accuracy, 97% precision, 100% recall, and 99% F1-score for the non-PCOS class and 100% precision, 94% recall, and 97% F1-score for the PCOS class. The evaluation of the machine-learning models revealed a confusion matrix with TP, TN, FP, and FN values of 37, 17, 1, and 0, respectively. The performance metrics, including precision, recall, F1 score, and accuracy, are listed in this table 2 and 3 for all criteria. Moreover, Figures 2–6 show the confusion matrix outputs for all machine learning classifiers considered in this work. Among the pretrained models, including Logistic Regression, SVM, Decision Tree, Random Forest, and XGBoost, the XGBoost model demonstrated the best performance in terms of recall, precision, and F1-score, with values of 0.1, 0.97, and 0.99, respectively. The complex nonlinear relationships among the clinical and biochemical variables of the PCOS dataset can be efficiently captured by XGBoost's gradient boosting technique. XGBoost L1 and L2 regularization prevents overfitting by prioritizing relevant features and minimizing noise in the PCOS dataset. The scale pos weight parameter addresses the class imbalance, effectively capturing patterns in underrepresented PCOS-positive cases. Meticulous tuning of XGBoost's hyperparameters, such as the learning rate, number of estimators, and maximum depth, significantly influ- ences model performance and sensitivity to subtle data patterns. The efficient convergence of XGBoost requires fewer iterations to achieve high performance and stability. Analyzing error types can highlight the precision of XGBoost in identifying PCOS patterns overlooked by other models, demonstrating its ability to capture essential data nuances. These advantages are reflected in the performance of the model, as evidenced by the highest accuracy score of 98% achieved using the XGBoost model. The accuracies of the other models ranged from 95% to 82%. The comparison graphs in Figures 7-11 show the highest and lowest values of the performance parameters obtained by the machine learning models. The XGBoost algorithm produced the highest values for all performance parameters, whereas the decision tree algorithm produced the lowest values. This discrepancy in performance demonstrates the superior capability of XGBoost to balance precision and generalization. The alignment between the training and testing accuracies further underscores the effectiveness of XGBoost, indicating minimal overfitting and robust generalization. XGBoost regularization prioritized relevant features while minimizing noise, and hyperparameter tuning ensured a stable performance. Future studies should test this model across diverse clinical populations for external validation and real-world applications. The effectiveness of the model was comprehensively assessed using precision, recall, F1-score, and AUC-ROC, particularly for the minority class, given the imbalanced nature of PCOS data. XGBoost's scale pos weight parameter addressed class imbalance, enhancing the model's ability to identify PCOS-positive cases, reducing false negatives, and improving recall. The learning curves demonstrate generalization performance and stability by showing an improvement in accuracy with increasing training samples. The achievement of 98% accuracy is notable for PCOS, for which symptom variability and diverse diagnostic criteria present challenges.

Table 2. Model performance on PCOS detection

Model's Name	Label	Precision	Recall	F1-Score	Accuracy
LR	No PCOS	0.93	0.10	0.96	95%
	PCOS	0.10	0.83	0.91	
SVM	No PCOS	0.95	0.10	0.97	96%
	PCOS	0.10	0.89	0.94	
Decision Tree	No PCOS	0.85	0.89	0.87	82%
	PCOS	0.75	0.67	0.71	
Random Forest	No PCOS	0.97	0.97	0.97	96%
	PCOS	0.94	0.94	0.94	
XGBoost	No PCOS	0.97	0.10	0.99	98%
	PCOS	0.10	0.94	0.97	

Table 3. Confusion matrix results for different machine learning models

Sr. No.	Machine Learning Model	True Positive	True Negative	False Positive	False Negative
1	Logistic Regression	15	37	0	3
2	SVM	16	37	0	2
3	Decision Tree	12	33	4	6
4	Random Forest	17	36	1	1
5	XGBoost	17	37	0	1

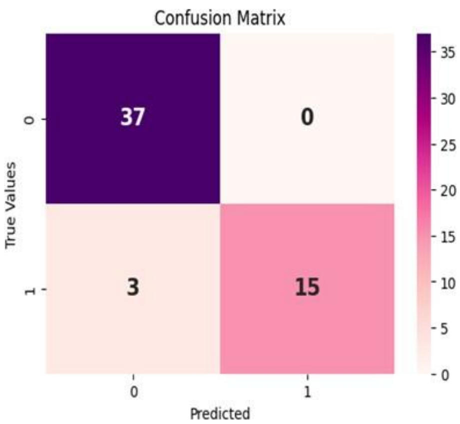

Fig.2. Confusion matrix for logistic regression model

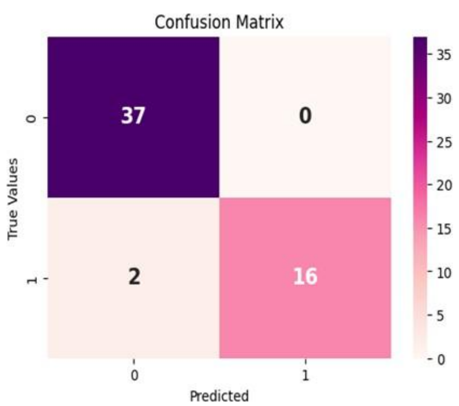

Fig.3. Confusion matrix for SVM model

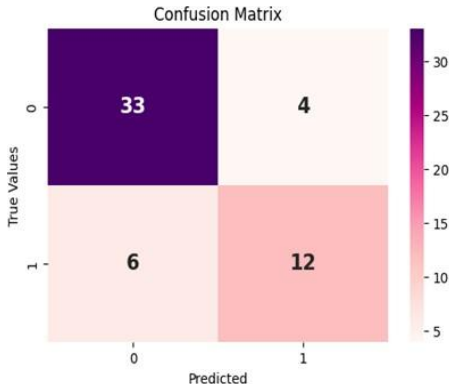

Fig.4. Confusion matrix for decision tree model

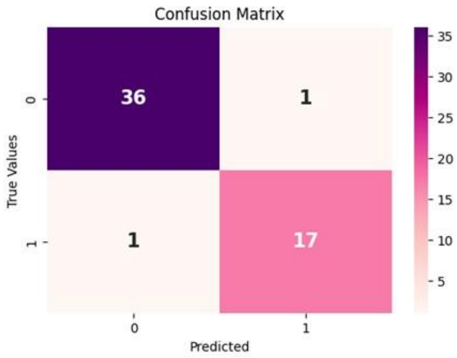

Fig.5. Confusion matrix for random forest model

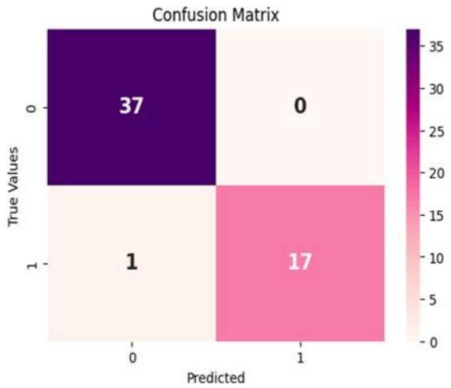

Fig.6. Confusion matrix for XGBoost model

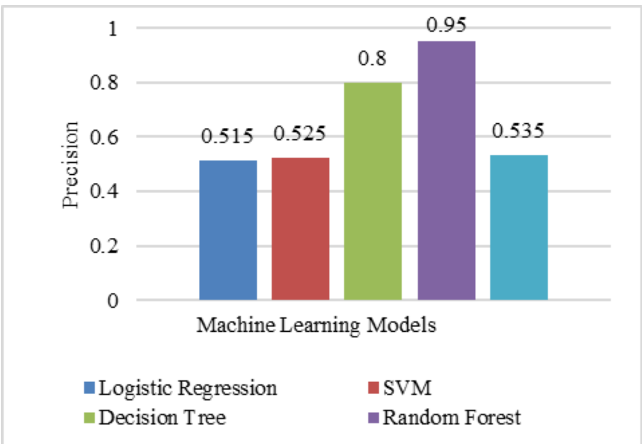

Fig.7. Precision comparison with different machine learning models

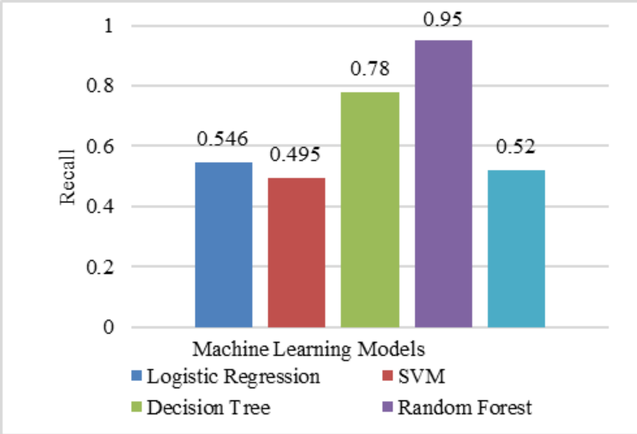

Fig.8. Recall comparison with different machine learning models

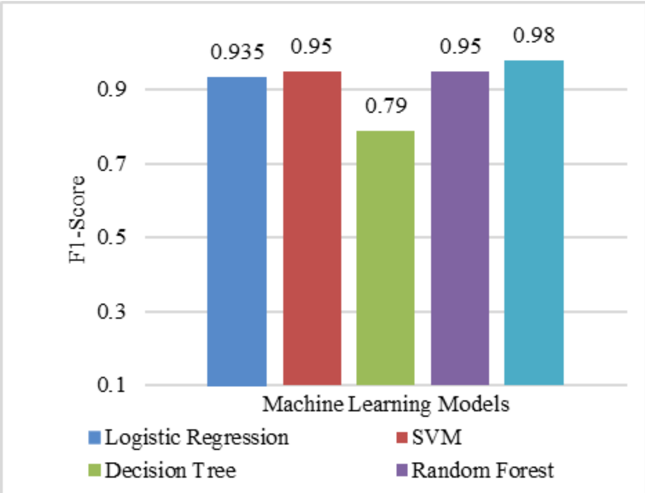

Fig.9. F1-Score comparison with different machine learning models

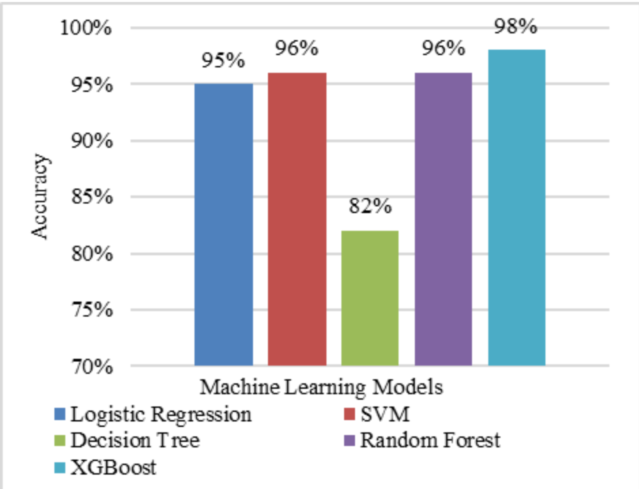

Fig.10. Accuracy comparison with different machine learning models

5.3. Implications of False Positives and False Negatives in Clinical Context

A false positive in a diagnostic test for polycystic ovary syndrome (PCOS) incorrectly indicates that an individual has the condition, leading to unnecessary treatments or interventions and potential adverse effects. For example, hormonal therapies or lifestyle changes based on an incorrect diagnosis can cause physical or psychological distress. Additionally, a false-positive diagnosis can induce anxiety and stress, affecting mental well-being and causing emotional distress related to concerns about infertility, weight gain, and other PCOS symptoms. False positives also increase healthcare costs due to unnecessary follow-up tests, consultations, and treatments, straining healthcare systems and resources. False negatives occur when a diagnostic test fails to identify an individual with PCOS, incorrectly suggesting the absence of the condition. Undiagnosed patients may miss timely interventions, allowing symptoms and complications such as infertility, diabetes, and cardiovascular disorders to progress. Early intervention is crucial for managing PCOS effectively, and without a proper diagnosis, patients may not adopt necessary lifestyle changes like dietary adjustments and exercise, which help manage symptoms and improve quality of life. Long-term health risks of untreated PCOS include metabolic syndrome, type 2 diabetes, and endometrial cancer, significantly impacting a patient's health trajectory. Precision, recall, and F1 score are crucial metrics for evaluating the ability of a model to reduce false positives and negatives. These indicators allow medical professionals to assess the efficacy of a model in practical clinical scenarios.

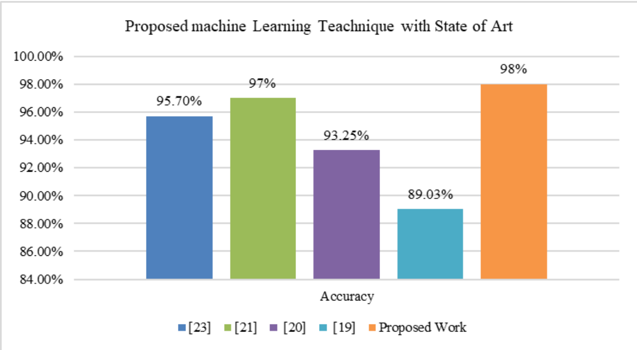

Fig.11. Performance of proposed XGBoost model with state of art technique

Table 4. Comparison of machine learning techniques for different datasets

Year of Publishing / Reference	Machine Learning Techniques	DataSet Description	Accuracy (%)
2023 / [23]	Adaptive Boosting, Extreme Gradient Boosting, RF, Categorical Boosting, Gradient Boosting	541	Gradient Boosting - 95.7%
2022 / [21]	EEMD, FFT, SVM	541	SVM - 97%
2022 / [20]	RF	541	RF - 93.25%
2022 / [19]	SVM with LR, DT, KNN, RF, ADABoost	541	SVM with LR - 89.03%
Proposed Work	XGBoost	541	98%

5.4. Performance Evaluation with State of Art Techniques

Table 4 demonstrates the most recent advancements in machine learning algorithms when applied to a single dataset. References, such as [23,21,20,19], and" Proposed Work (XGBoost)," each analyze multiple machine-learning approaches and report their respective accuracies on the dataset. Reference [23] achieves 95.70% accuracy in classification tasks by combining boosting algorithms, such as Adaptive Boosting and XGBoost. Additionally, it achieves 97% accuracy using signal processing techniques, such as EEMD and FFT with SVM, demonstrating effective feature extraction methods. Reference [20] focuses on RF and achieved 93.25% accuracy, demonstrating the effectiveness of ensemble learning methods for classification. Reference [19] investigated various classifiers with SVM paired with LR, obtaining 89.03% accuracy and demonstrating the value of hybrid techniques in enhancing the classification outcomes. The" Proposed Work" stands out with the highest reported accuracy of 98% on the dataset, as shown in figure 6.

6. Conclusions

This research evaluates various machine-learning models for predicting Polycystic Ovary Syndrome, emphasizing their performance on numerical data. Traditional models such as Logistic Regression and Support Vector Machine achieved accuracies of 95% and 96%, respectively, while the ensemble model XGBoost proved to be the most effective with 98% accuracy. This highlights the importance of advanced algorithms for managing the complexity of PCOS data and enhancing diagnostic precision. Moreover, this study can be strengthened by acknowledging its limitations, particularly the impact of data imbalance on model performance. Future research should explore techniques to address these issues, such as additional sampling methods or alternative metrics that reflect the clinical significance of PCOS diagnosis. The findings of this study suggest several directions for future research, including the integration of diverse datasets with demographic and lifestyle factors to enhance model accuracy and generalizability. Exploring hybrid models that combine different machine learning techniques may further improve predictive capabilities. Addressing these areas can build upon this foundational work, advance AI applications in healthcare, and improve outcomes in women with PCOS.

Acknowledgment

The authors would like to thank Chitkara University Research and Innovation Network, Chitkara University, Punjab, India for their support.

References

[1] S. Viswanathan, R. Jiji, B. C. Nayana and C. Baby, "Pregnancy complications associated with polycystic ovary syndrome: A cross-sectional study," World J Pharm Res, 11, 2022, 1539-1552.

[2] L.H. Zeng, S. Rana, L. Hussain, M. Asif, M. H. Mehmood, I. Imran, and S.N. Abed, "Polycystic ovary syndrome: a disorder of reproductive age, its pathogenesis, and a discussion on the emerging role of herbal remedies," Frontiers in Pharmacology,13, 2022, https://doi.org/10.3389/fphar.2022.874914874914- 874936.

[3] S. A. Bhat, "Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms (Doctoral dissertation, Dublin, National College of Ireland)," 2021, 1-30.

[4] A. A. Choudhury, and V.D. Rajeswari, "Gestational diabetes mellitus-a metabolic and reproductive disor- der," Biomedicine & Pharmacotherapy, 143, 2021, 112183-112201.

[5] A. Denny, A. Raj, A. Ashok, C. M. Ram, and R. George, "I-HOPE: Detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques," in TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019.

[6] V. Thakre, "PCOcare: PCOS detection and prediction using machine learning algorithms," Biosci. Biotechnol. Res. Commun., vol. 13, no. 14, pp. 240–244, 2020.

[7] S. S. Deshpande and A. Wakankar, "Automated detection of polycystic ovarian syndrome using follicle recognition," in 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, 2014.

[8] S. Ahmed, M. S. Rahman, I, Jahan, M.S. Kaiser, A. S., Hosen, D. Ghirime, and S. H. Kim, "A review on the detection techniques of polycystic ovary syndrome using machine learning," IEEE Access, vol. 11, pp. 86522–86543, 2023.

[9] D. Hdaib, N. Almajali, H. Alquran, W. A. Mustafa, W. Al-Azzawi, and A. Alkhayyat, "Detection of polycystic ovary syndrome (PCOS) using machine learning algorithms," in 2022 5th International Conference on Engineering Technology and its Applications (IICETA), 2022.

[10] R. Kaur, R. Kumar, and M. Gupta, "Lifestyle and Dietary Management Associated with Chronic Diseases in Women Using Deep Learning. Combating Women's Health Issues with Machine Learning," pp. 59–73, 2024.

[11] I. B. CICEK, Z. KUCUKAKCALI, and F. H. YAGIN, "Detection of risk factors of PCOS patients with Local Interpretable Model-agnostic Explanations (LIME) Method that an explainable artificial intelligence model," The Journal of Cognitive Systems, 6, 2021, 59-63.

[12] Neto, C., Silva, M., Fernandes, M., Ferreira, D., and Machado, J., Prediction models for Polycystic Ovary Syndrome using data mining. In International Conference on Advances in Digital Science, 2021, 210-221.

[13] Rathod, Y., Komare, A., Ajgaonkar, R., Chindarkar, S., Nagare, G., Punjabi, N., & Karpate, Y. (2022, July). Pre- dictive Analysis of Polycystic Ovarian Syndrome using CatBoost Algorithm. In 2022 IEEE Region 10 Symposium (TENSYMP) 2022, 1-6.

[14] M. Marreiros, D. Ferreira, C. Neto, D. Witarsyah, and J. Machado, "Classification of Polycystic Ovary Syndrome Based on Correlation Weight Using Machine Learning," In Big Data Analytics and Artificial Intelligence in the Healthcare Industry, IGI Global, 2022, 150-176.

[15] H. Danaei Mehr, and H. Polat, "Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques," Health and Technology, 12, 2022,137-150.

[16] P. Bhardwaj, and P. Tiwari, "Manoeuvre of Machine Learning Algorithms in Healthcare Sector with Application to Polycystic Ovarian Syndrome Diagnosis," In Proceedings of Academia-Industry Consortium for Data Science, Springer, Singapore, 2022, 71-84.

[17] G. Sinthia, T. Poovizhi, and R. Khilar, "Analysis on Polycystic Ovarian Syndrome and Comparative Study of Differ- ent Machine Learning Algorithms," In Advances in Intelligent Computing and Communication, Springer, Singapore, 2022, 191-196.

[18] R. Subha, B. R. Nayana, R. Radhakrishnan, and P. Sumalatha,"Computerized Diagnosis of Polycystic Ovary Syn- drome Using Machine Learning and Swarm Intelligence Techniques;" 2022, 3330-3357.

[19] K. Rakshitha, and N. Naveen, "Op-RMSprop (Optimized-Root Mean Square Propagation) Classification for Predic- tion of Polycystic Ovary Syndrome (PCOS) using Hybrid Machine Learning Technique," International Journal of Advanced Computer Science and Applications,13, 2022, 0130671-0130681.

[20] S.Tiwari, L. Kane, D. Koundal, A. Jain, A. Alhudhaif, K. Polat, ... and S. A.Althubiti, "SPOSDS: A Smart Polycystic Ovary Syndrome Diagnostic System Using Machine Learning," Expert Systems with Applications, 2022, 117592- 117622.

[21] Tempelman, J. R., Wachtor, A. J., Flynn, E. B., Depond, P. J., Forien, J. B., Guss, G. M., ... & Matthews, M. J. Detection of keyhole pore formations in laser powder-bed fusion using acoustic process monitoring measure- ments. Additive Manufacturing, 55, 2022, 102735-102751.

[22] Khanna, V. V., Chadaga, K., Sampathila, N., Prabhu, S., Bhandage, V., & Hegde, G. K. A distinctive explainable machine learning framework for detection of polycystic ovary syndrome. Applied System Innovation, 6(2), 2023, 1-32.

[23] Suha, S. A., & Islam, M. N. Exploring the dominant features and data-driven detection of polycystic ovary syndrome through modified stacking ensemble machine learning technique. Heliyon, 9(3), 2023, 1-21.

[24] Svm-PCOS diagnosis. Retrieved, from Kaggle.com website: https://www.kaggle.com/code/swabbie8/svm-pcos-diagnosishttps://www.kaggle.com/code/swabbie8/svm-pcos-diagnosis, 2023.

[25] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[26] Wang, C., Deng, C. and Wang, S., 2020. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern recognition letters, 136, pp.190-197.
[27] Rodriguez, J.J., Kuncheva, L.I. and Alonso, C.J., 2006. Rotation forest: A new classifier ensemble method. IEEE transactions on pattern analysis and machine intelligence, 28(10), pp.1619-1630.
[28] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
[29] Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, pp.1189- 1232.
[30] Maimon, O. and Rokach, L. eds., 2005. Data mining and knowledge discovery handbook (Vol. 2, No. 2005). New York: Springer.
[31] Terven, J., Cordova-Esparza, D.M., Ramirez-Pedraza, A., Chavez-Urbiola, E.A. and Romero-Gonzalez, J.A., 2023. Loss functions and metrics in deep learning. arXiv preprint arXiv:2307.02694.

Authors' Profiles

Shikha Prasher completed her B. Tech in Computer Science and Engineering from Om Institute of Engineering and Technology, Hisar in 2014 and M.E. in Computer Science and Engineering from Chitkara University Research and Innovation Network, Chitkara University, Punjab in 2024. She has four years of teaching experience at various engineering colleges. She has published one patent and 30 Scopus-indexed international conference papers. Her current research interests focus on the fields of Machine Learning and Deep Learning.

Dr. Leema Nelson is an Associate Professor in Research at Chitkara University Research and Innovation Network (CURIN), specializing in advanced research. She has completed her M.E Computer Science and Engineering (specialization in Knowledge Engineering) and Ph.D. in Computer Science and Engineering from College of Engineering Guindy, Anna University Chennai in 2018. She has seven years of research and eleven years of teaching experience at various engineering colleges. She has received college first-rank award for her graduate study. She selected one of the 25 doctoral candidates from all over India for practical-oriented training for new researchers in India in 2023 by the University of Cologne, Germany. She has six patent grants and ten published patents. Moreover, she has published 87 journals/international conferences, including SCI and Scopus Indexed. She has held multiple roles at various conferences, including serving as a Session Chair and Advisory Board Member. Her current research interests are concentrated in the fields of Machine Learning, Deep Learning, and Federated Learning.

Manal Gafar received B.Sc. degree from the networks and communication program, Faculty of Electronic Engineering, Menouf, Menoufia University, Egypt, in 2023 with a GPA of 3.64 out of 4, which equates to 91.43%. She is currently a Teaching Assistant in the Cyber Security and Networks Program at the University of East London, European Universities in Egypt (EUE), located in The New Administrative Capital, Egypt. She has actively participated in various training courses focused on computer networks, communications, and security, including Computer Networks Internship at ITI, CCNA 200–301 at NTI, CCNP ENCOR at NTI, CCNP ENARSI at NTI, and Network and Information Security NIS at NTI. Manal was certified by Huawei in the fields of cloud computing, routing, and switching. Manal has successfully completed an advanced diploma in Engineering Sciences (Master Pre-Courses) during the calendar year 2023/2024 with an outstanding degree, and is currently enrolled in a master's program focusing on Software-Defined Networking (SDN) and cybersecurity. Manal's research interests include cyber/network security, Internet of Things (IoT), and Machine Learning.

How to cite this paper: Shikha Prasher, Leema Nelson, Manal Gafar, "NIPP: Non-Invasive PCOS Prediction using XG-boost Machine Learning Model", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.1, pp.82-95, 2025. DOI:10.5815/ijitcs.2025.01.06