RESEARCH PAPER

AI-DRIVEN PCOS PREDICTION USING A NEUROENDOCRINE STRESS INDEX AND WEARABLE INSPIRED FEATURES

[1]K R Revathy, [2]S T Padmapriya, [3]D Meenakshi
[1,2,3]Applied Mathematics and Computational Science,
Thiagarajar College of Engineering, Madurai, Tamilnadu, India
[1] revathykr@res.tce.edu, [2]stpca@tce.edu, [3]meenakshid@res.tce.edu,
Corresponding Author: stpca@tce.edu

Abstract
Polycystic ovary syndrome (PCOS) is a prevalent endocrine disorder in women of reproductive age, yet early risk stratification remains challenging due to its multifactorial nature. This paper proposes an AI-driven wearable framework for early PCOS prediction built around a novel Neuroendocrine Stress Index (NSI) — a composite biomarker aggregating pulse rate, thyroid stimulating hormone, prolactin, random blood sugar, and BMI — which captures the cumulative physiological stress burden associated with PCOS onset. The framework integrates clinical, hormonal, metabolic, and wearable-derived signals (heart rate, sleep duration, physical activity, stress level) from the Kaggle PCOS dataset, processed through a structured pipeline of median imputation, standardization, and feature selection. An ensemble of Random Forest, Logistic Regression, and XGBoost classifiers is trained and evaluated, achieving up to 83.4% accuracy. SHAP-based explainability identifies follicle count, AMH, BMI, and the NSI as the most influential predictors, enhancing clinical transparency. Results confirm that embedding stress-related wearable signals alongside conventional clinical features improves prediction accuracy and supports personalized, interpretable healthcare decision-making for women at risk of PCOS.
Keywords: Artificial intelligence, Biomarkers, Body mass index, Clinical decision support-tools, Data preprocessing, Early disease prediction, Explainable artificial intelligence, Feature selection, Hormonal imbalance, Machine learning, Neuroendocrine stress index, Personalized healthcare, Polycystic ovary syndrome, Predictive modeling, Random forest classifier, SHAP (SHapley Additive exPlanations), Stress biomarkers, Wearable sensing devices, Women's health, XGBoost.

How to cite this article: Revathy KR, Padmapriya ST, Meenakshi D. AI-Driven PCOS Prediction Using a Neuroendocrine Stress Index and Wearable Inspired Features. Int J Drug Deliv Technol. 2026;16(62s): 429-442. DOI: 10.25258/ijddt.16.62s.50

Source of support: Nil.

Conflict of interest: None.

1. INTRODUCTION

Polycystic ovary syndrome (PCOS) is an endocrine-metabolic complication that is a complication of the complex, and thus, this complication is one of the most highly observed health conditions among women of childbearing age in the entire world. PCOS has been characterized by hormonal imbalances, menstrual cycle disorders, polycystic ovaries, and associated metabolic problems like obesity and insulin resistance. Diagnosis of PCOS is important at early stages since its delay can result in serious health problems like infertility, diabetes, and heart diseases. However, PCOS diagnosis is often complicated by the presence of varied symptoms and the involvement of numerous clinical and physiological parameters.

In the last few years, data-driven machine learning methods have been extensively investigated in the field of healthcare for disease prediction and decision-making in the process of disease diagnosis. Such models can analyze large volumes of clinical and laboratory data and reveal non-trivial trends that are hard to discover with the help of traditional statistical tools. In the case of the disease PCOS, several studies have employed Gradient boosting, logistic regression, and random forest machine learning models to make predictions on the disease using clinical and hormonal factors. However, the majority of the studies have focused on clinical data alone and have not considered other factors like lifestyle and physiology that might affect the disease.

We can now have our physiology continually tracked with new wearable technology such as our heart rate, sleep patterns, stress levels, and activities. Our lifestyle is very closely related to hormonal balance and metabolic health, and both of these factors have major roles to play in PCOS. We can improve the accuracy of prediction and gain better insights into PCOS through this.

Moving into the future, wearables will have the ability to go beyond sensors and be emotionally intelligent companions that can react to our emotions and deeds. For instance, through the application of AI-powered adaptive interfaces, wearables could adapt in real-time to respond to our levels of stress, sleep patterns, and levels of physical activities based on sensor readings. In this paper, we outline a PCOS prediction framework that can be used to power emotionally intelligent wearables for women's health.

In the paper, A new Neuroendocrine Stress Index is also proposed for the detection of stress-related physiological signals, which may also play a part in the prediction of PCOS. Different machine learning approaches are used to evaluate the best approach for the prediction of PCOS, including an ensemble method.

The paper structure is as follows: This paper has eight sections. Section 3 addresses the data, both the conventional data and the wearable data, and the selection Section 2 reviews the literature, with particular reference to how machine learning has been applied to detect PCOS and the most significant gaps of the literature. Section 4 discusses the proposed approach, including the machine learning models, ensemble models, and the Neuroendocrine Stress Index. Section 5 covers implementation of models and the measures of evaluation. Section 6 has been anchored on the interpretation of the experimental findings, relevance of features, and the model explanation through SHAP methodology. Section 7 deals with the study relative to the literature and Section 8 is a conclusion of the paper where future work is mentioned.

2. RELATED WORKS

Over the last several years, Various studies have been conducted to pursue various directions, including traditional machine learning strategies to the most modern deep learning models, and clinical decision support systems and explainable AI**.** In this paper, we will discuss the major research works done on the prediction of PCOS.

2.1 Machine Learning–Based PCOS Prediction

Many research works have employed machine learning techniques to predict PCOS based on clinical and hormonal characteristics. For instance, Ramteke and Raut developed an interpretable AI system that utilized different techniques for feature selection, including Sequential Forward Selection and Boruta. The proposed method showed better prediction accuracy and interpretability through SHAP and LIME analysis [1].

Rahman et al. have developed an online machine learning platform for the early detection of PCOS. They have used different machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, AdaBoost, and SVM, and Mutual Information for feature selection to make the models more efficient. Random Forest was the most accurate of the models.

In a comparative study, Yadav et al. has analyzed the performance of numerous machine learning models in the diagnosis of PCOS. The results showed the superiority of recent classification models over traditional statistical models in terms of accuracy in detecting PCOS cases [4]. This shows the increasing trend in the application of machine learning to deal with complicated medical information to provide better diagnostic support in women's health. Further, in the study by Mamatha and Sravani, Random Forest and XGBoost models were implemented to diagnose PCOS, yielding good classification accuracy, which shows the potential of ensemble learning models in dealing with medical information.

Rani et al. took this research further and used machine learning to assess hormonal and metabolic markers. Some of the predictive values of different variables including Anti-Mullerian Hormone (AMH), Body Mass Index (BMI) and follicle counts were also highlighted in this study as a predictor of PCOS. Rao and Ramesh also used predictive analytics techniques to analyze various factors related to PCOS and develop predictive models to help in early detection.

2.2 Ensemble Learning and Optimization Techniques

Ensemble learning methods have been found to give significant improvements in PCOS prediction performance. Kumar and Sharma proposed an ensemble model for PCOS detection using Random Forest and XGBoost algorithms. The model proposed was more accurate in PCOS classification. To maximize the algorithmic performance and to enhance the prediction, various machine learning models were developed. The prediction performance of the proposed model was better than that of other models. The suggested model is one of the ways optimization can be conducted to enhance the performance of machine learning models in the diagnosis of women health. [14]

Prasher et al. have also proposed a non-invasive PCOS prediction model using the XGBoost algorithm for detection using numerical health data. Their results have also proved that boosting algorithms can be used for improving accuracy in health care [13].

2.3 Deep Learning Approaches for PCOS Detection

Deep learning techniques have also been used to develop models for PCOS prediction. Islam et al. introduced neural network-based techniques for healthcare dataset analysis. This illustrated that neural network models are capable of effective identification of complicated trends in the PCOS datasets [7]. In addition, Ahmad et al. introduced dataset balancing using SMOTE and CNN, LSTM, and hybrid CNN-LSTM models for PCOS prediction. This demonstrated that deep learning models can achieve accuracy and enhance classification performance [8].

2.4 Healthcare Systems and Explainable AI in PCOS Prediction

The current research studies have focused on developing smart health platforms for the detection and diagnosis of PCOS. Gupta et al. proposed a clinical decision support framework for the diagnosis of PCOS by healthcare professionals using machine learning algorithms and medical data of patients [10]. Jaganathan and Natesan proposed a blockchain-based explainable artificial intelligence framework for the detection of PCOS, providing secure data sharing in the healthcare domain [3].

Another area that has been explored is large-scale healthcare data analysis. The predictive models that Zad et al. created to identify PCOS utilized the electronic health records to effectively analyze real-life healthcare data, demonstrating that machine learning algorithms can be applied to analyze the healthcare system. Systematic reviews have also identified that artificial intelligence plays a significant role in enhancing PCOS diagnosis.

2.5 Research Gap

While considerable advancements have been made by prior research in the prediction of PCOS using various approaches, most of the existing methods are based on clinical, hormonal, and metabolic factors. However, not many research works have considered the incorporation of physiological stress-related biomarkers in their predictive models. In addition, not many research works have considered combining stress-related biomarkers and ensemble machine learning approaches.

In this context, this paper proposes a novel model that can address existing limitations and challenges. This model incorporates various clinical, hormonal, and metabolic aspects with a novel Neuroendocrine Stress Index. This index consists of a range of stress indicators (pulse rate, thyroid-stimulating hormone, prolactin, blood sugar level, and body mass index).

Moreover, this paper uses ensemble techniques that combine Logistic Regression, Random Forest, and XGBoost algorithms. This is to improve both accuracy and stability. This framework is expected to improve early detection of PCOS and better understand the relationship between stress and PCOS.

3. MATERIALS AND METRICS

The section below briefly describes the dataset that has been used in this research, along with the evaluation measures that are considered to check how accurately the machine learning models are able to predict PCOS. Various attributes are considered in this dataset that can be used to identify patterns that are related to Polycystic Ovary Syndrome (PCOS). In addition to this, evaluation measures are considered to assess the accuracy and reliability of the predictive models.

3.1 Dataset Description

A dataset that was acquired as a result of the Kaggle PCOS competition is used in the paper. This dataset contains medical records of women with various attributes that are both clinical and physiological in nature. This dataset has various attributes such as hormonal levels, metabolic factors, ultrasound scans, and health status. The data set includes a number of attributes and they can be categorized into three broad categories:

3.1.1. Clinical Features

- Body Mass Index (BMI)
- Follicle count in the left ovary (Follicle No. (L))
- Follicle count of the right ovary (Follicle No. (R))
- Endometrium thickness
- Hormone presence (Follicle Stimulating Hormone (FSH), Anti-Mullerian Hormone (AMH) and Luteinizing Hormone (LH)).

3.1.2. Metabolic Features

- Random Blood Sugar level
- Waist-Hip Ratio
- Weight and metabolic indices associated with insulin resistance.

3.1.3. Wearable-Derived Features

- Heart rate
- Sleep duration
- Physical activity level
- Stress level

In addition to the above features, a Neuroendocrine Stress Index was proposed as a part of the feature engineering process to account for the impact of physiological stress as well as hormonal imbalance on the risk of developing PCOS. Several machine learning models were then trained using the preprocessed data. As a part of the pre-processing step, the missing data in the dataset were handled, the feature values were standardized, and the most informative features were identified.

3.2 Evaluation Metrics

To understand how effectively these models predict PCOS, various measures for model evaluation were used. These, according to the confusion matrix, are the measures of how well the model classifies positive and negative cases. There are four elements of the confusion matrix:

True Positive (TP): The cases when PCOS is rightfully detected.

True Negative (TN): The cases of the correct identification of non-PCOS.

False Positive (FP): The instances where non-PCOS is wrongly identified as PCOS False Negative (FN): The instances where PCOS is wrongly identified as non-PCOS

Different measures may be obtained using these components.

3.2.1. Accuracy

Accuracy is also used to determine the percentage of the predictions that the model has made accurately and thus it is used to measure the ability of the model to distinguish between the two categories of women, namely, the ones suffering from PCOS and the ones not suffering from the condition.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(1)}$$

The higher the accuracy of the model, the higher the ability of the model to differentiate between the two categories of women, namely, the ones suffering from PCOS and the ones not suffering from the condition. However, it has also been seen that the accuracy of the model may not always be a true reflection of the ability of the model to differentiate between the two categories of women, namely, the ones suffering from PCOS and the ones not suffering from the condition, in cases where the dataset may be unbalanced in nature.

3.2.2. Precision

Precision is used to determine the amount of the real number of positive cases out of the cases that the classifier has classified as positive. Precision is computed as:

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(2)}$$

This measure is particularly useful when the cost of FP is high because it highlights the importance of positive predictions.

3.2.3 Recall (Sensitivity)

Also recall or sensitivity measures the capacity of the model to effectively recognize true PCOS cases. Recall is given by:

$$Recall = \frac{TP}{\mathrm{TP + FN}} \quad \textbf{(3)}$$

A high recall indicates that real PCOS cases are being identified by the model, thus reducing the chances of missing real PCOS patients who need to be taken care of.

3.2.4. F1-Score

F1 Score is a statistics used to combine both the precision and recall of the data by taking the harmonic mean:

$$\mathrm{F1-Score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{Precision + Recall} \quad \textbf{(4)}$$

F1 Score is a trade off between the three and therefore it is applicable in situations where false works. Positives and false negatives are harsh repercussions..

3.2.5. Receiver Operating Characteristic – Area Under Curve (ROC–AUC)

The Receiver Operating Characteristic (ROC) curve is also given in the form of a plot of the false positive rate versus the true positive rate when the threshold is varied under different settings. The Area Under the Curve (AUC) determines the overall performance of the model when it comes to distinguishing between the PCOS and non-PCOS cases. The value of AUC is large, which suggests that it has high discriminative power.

$$\mathrm{ROC-AUC} = \int_0^1 TPR \; d(FPR) \quad \textbf{(5)}$$

Where:

$$TPR = \frac{TP}{TP + FN} \quad \textbf{(6)}$$

$$FPR = \frac{FP}{FP + TN} \quad \textbf{(7)}$$

The larger the ROC-AUC score, the larger the ability of the model to differentiate between PCOS and non-PCOS classes with different threshold values.

3.3. METHODOLOGY

The paper proposes a machine learning framework that can potentially aid in the early detection of PCOS. The model takes into account several types of input data, such as clinical data, hormonal levels, lifestyle traits, and wearable devices. In addition to traditional medical characteristics, this framework incorporates a Neuroendocrine Stress Index that incorporates various physiological and lifestyle characteristics such as heart rates, sleep patterns, stress levels, and physical activities that can potentially reflect the role of lifestyle and neuroendocrine stress in the causation of PCOS. The general structure of this framework is presented as Figure 1. Such a predictive model can even be used together with emotionally intelligent wearables that might implement AI-based adaptive interfaces.

.

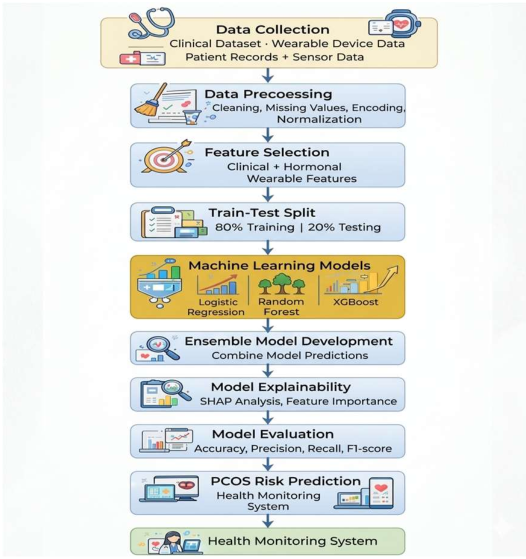

Figure1. Framework for PCOS prediction that integrates ensemble machine learning techniques with SHAP-based interpretability

3.3.1. Data Collection

The dataset used in this research consists of information regarding various clinical parameters from women who have undergone testing for Polycystic Ovary Syndrome (PCOS). The dataset was provided in CSV format and has 541 instances. The dataset has various physiological, hormonal, and lifestyle-related parameters for each patient. The dataset has various important parameters such as Demographic parameters (Age, Weight, Height), Metabolic parameters (BMI, Random Blood Sugar), Hormonal parameters (TSH, PRL, AMH, LH, FSH), Ultrasound parameters (Follicle count, Endometrium), Lifestyle parameters (Exercise, Fast food). The dataset has PCOS diagnosis as one of its parameters, which is in binary form (Yes/No). The parameter PCOS diagnosis represents a notion that a person in the dataset was diagnosed with PCOS clinically.. The dataset has various biological and physiological parameters, making it a comprehensive overview of various factors associated with PCOS.

3.3.2. Data Preprocessing

A preprocessing is a significant part of data integrity and the effectiveness of machine learning models maximization. The data set was improved by a variety of preprocessing methods to make it fit into the further analysis. First, missing data in the data set was identified and replaced using appropriate statistical methods. Outliers in critical data, especially in clinical and metabolic data, were also identified and handled appropriately. Further, data standardization was performed for numeric data, while for categorical data, appropriate coding was done. In addition, irrelevant columns in the data set, such as identifier columns and duplicate data, were removed to ensure minimal variation in data and improve the efficiency of the model in generalizing effectively.

Handling of Missing Values

There were a few instances where certain values were inconsistent within the dataset. Median imputation was used to deal with inconsistencies. In this method, missing values are filled up with medians of their feature values.. This technique is used because it is not heavily impacted by extreme values and maintains the distribution of values.

Data Type Conversion

There were a few instances where certain values were stored in string format instead of appropriate data types. These values were changed to appropriate data types to enable mathematical calculations. Additionally, certain categorical values such as “Yes/No” or “Y/N” were changed to appropriate numeric values. These values were changed to either 0 or 1. In addition, date and time values were changed to appropriate datetime values. In certain instances, date and time values were used to calculate derived feature values such as age. In this conversion process, the process of any values being found invalid was treated with care so as to avoid any errors that may arise when it runs.

Feature Cleaning

The irrelevant columns, such as serial numbers and patient file numbers, were eliminated since they do not contribute to the prediction of PCOS. Elimination of these attributes ensures that there is no bias in the data. This phase enhances computational power and the algorithms are then able to concentrate on the aspects that are of actual use. Thus, the cleaned dataset is a collection of significant medical, metabolic, and wearable-derived features necessary for PCOS prediction.

Integration of Wearable Data

Besides the clinical features, wearable-derived features (heart rate, stress level, sleep duration, and physical activity) were also added to the dataset. These characteristics provide additional information about physiological conditions, which provide a more detailed picture of lifestyle-related factors that can cause PCOS. With wearable-derived features incorporated, the model can find out the nuanced relationships and underlying structures concerning hormonal imbalance and stress.

Feature Scaling

To ensure that all features are of equal importance in the model, numerical features were standardized by applying the StandardScaler technique. This makes sure that all the features are on standard scale with a mean of zero and unit variance.

3.3.3. Feature Engineering: Development of Neuroendocrine Stress Index

One of the important contributions of the present work is the development of a Neuroendocrine Stress Index (NESI) that aims to quantify the stress

levels in the patients. PCOS has a strong link with hormonal imbalances, metabolic problems, and stress-related biological mechanisms. In view of the above, a stress index was formulated that would represent the combined effect of the different stress-related biomarkers in the patients. Based on the medical relevance of the different features, the Pulse Rate (in bpm), Thyroid Stimulating Hormone (TSH), Prolactin (PRL), Body Mass Index (BMI), and Random Blood Sugar (RBS) were selected as the different features of the stress index. Each of the features was standardized by applying the z-normalization technique. Then, the Neuroendocrine Stress Index was calculated by taking the mean of the standardized values of the different features.
Mathematically,

$$\text{Neuroendocrine Stress Index} = \frac{(Z_{Pulse} + Z_{TSH} + Z_{PRL} + Z_{RBS} + Z_{BMI})}{5} \quad \textbf{(8)}$$

where Z denotes the standardized value of the features. This index assists in the accumulation of the effect of stress-related biological factors in the development of PCOS

3.3.4. Feature Selection

Feature selection is a vital step in the machine learning process that allows the identification of the features that have the most significant impact on the prediction outcomes related to PCOS. In the current study, the selection of the most important features was performed after the preprocessing and feature engineering steps of the dataset. Wearable and clinical features were considered in the analysis in order to determine their impact on the detection of PCOS.

The features considered in the research for prediction were divided into three groups: clinical features such as Body Mass Index (BMI), the number of follicles in the left and right ovaries, hormone levels, and endometrial thickness; metabolic features such as random blood sugar and waist/hip ratio; Additionally, derived features such as heart rate, sleep, stress, and activity derived from wearables are included. The combination of all the above features has the purpose of covering medical, metabolic, and lifestyle factors In addition, the Neuroendocrine Stress Index combined was incorporated as a constructed attribute to account for the effect of lifestyle and neuroendocrine factors in the risk of PCOS.

The correlation coefficient that is utilized to define the association between two attributes is presented by

$$r = \frac{\Sigma\,(\mathrm{X} - \bar{\mathrm{X}})(\mathrm{Y} - \bar{\mathrm{Y}})}{\Sigma\,(\mathrm{X} - \bar{\mathrm{X}})^2\,\Sigma\,(\mathrm{Y} - \bar{\mathrm{Y}})^2} \quad \textbf{(9)}$$

In the above equation, the independent attributes X and the dependent attribute Y represent the selected attributes and the diagnosis of PCOS, respectively. These selected attributes were then employed in the training of the ensemble machine learning models.

3.3.5. Train–Test Split

The identification of the respective features was followed by the creation of a training set and testing set to confirm the model performance. The training set was used to train the machine learning algorithms, while the testing set was used to verify the accuracy of the predictions made by the model. This ensures that the model has the ability to perform well on unseen cases and prevents the possibility of over-fitting the model.

3.3.6. Machine Learning Model Development

To ensure that a reliable PCOS prediction model was created, a number of machine learning algorithms were used and compared in the development of the model. These algorithms were trained using the features that were extracted in the pre-processing and feature selection steps. Logistic Regression, Random Forest, and XGBoost were used in the development of the PCOS prediction model due to their ability to effectively work with medical datasets and their ability to identify the intricate relationship that exists between the variables in the dataset. Logistic Regression, a popular statistical technique used in binary classification problems, uses the logistic function to determine the probability of the occurrence of PCOS:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \quad \textbf{(10)}$$

where X is an input features, and β is amodel coefficient.

Random Forest is a group learning algorithm. The model has many decision trees which are integrated to increase the precision of the decisions. The projections of various decision trees make the model more robust.. The model is less likely to overfit, which is an essential factor in medical data.

$$F(x) = N \sum_{i=1}^{N} T_{i(x)} \quad \textbf{(11)}$$

In this representation, Ti(x) represents the prediction from each decision tree in the model, and N represents the total decision trees in the model.

Another boosting algorithm is XGBoost. It is because the model enhances accuracy in prediction through regularization to prevent overfitting.
XGBoost objective purpose is.

$$Obj = \sum_{i=1}^{n} l(y_i, \tilde{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \quad \textbf{(12)}$$

In this expression, l refers to the loss function, and Ω specifies the regularization term.

3.3.7. Ensemble Model Development

In order to improve the stability of this system and increase its predictive capability, ensemble learning was used. Ensemble learning is a method in which a group of models is applied together in order to make a single and more accurate prediction. In this case, a model suite comprising of the Logistic Regression model, the Random Forest model, and the XGBoost model was applied to produce one and only reliable prediction of the risk of PCOS.. The ensemble model works in the following manner:

1. The models are first trained on the training data.
2. Once the models are trained, each of them is used to make a separate prediction for a given input.
3. The model then average or vote on their prediction.
4. The final prediction is then made by combining all the models.

3.3.8. Model Explainability

To increase transparency and interpretability of the developed machine learning-based approach, explainable AI techniques have been adopted. In this context, SHAP analysis has been conducted to understand the contributions of different features and identify the most impactful factors that influence the prediction of PCOS. The SHAP analysis may be performed to attain:

- To understand the influence of different features that are considered for the prediction of PCOS by the developed machine learning-based approach
- To identify critical factors that are associated with PCOS
- To increase confidence in the developed machine learning-based approach for the prediction of PCOS

The developed approach demonstrates a comprehensive and effective integration of clinical records and hormonal profiles and different features extracted from wearable devices for the prediction of PCOS. In this context, lifestyle and physiological factors like heart rate, sleep patterns, perceived stress levels, and physical activity are considered for the prediction of PCOS. These factors can increase the efficiency of machine learning by identifying the patients who are at higher risk of developing PCOS. In addition, different algorithms are considered in combination with the ensemble technique to increase the accuracy of the developed approach.

4. RESULTS AND DISCUSSION

4.1.1 Dataset Analysis and Preprocessing Results

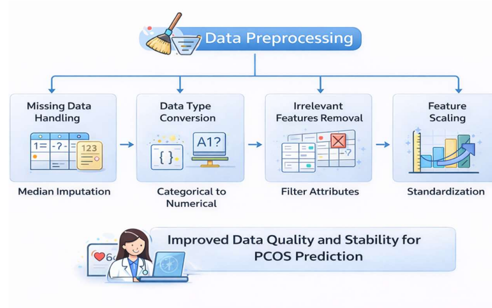

Figure 2: Data Preprocessing Pipeline for PCOS Prediction

In the picture above, you can see an example of a machine learning preprocessing process pipeline. This includes the following processes to reduce missing data; imputation (replacing missing values with the median) and converting categorical attributes into numeric values (also known as dummy/indicator variables), removing all irrelevant variables and applying feature scaling by way of standardizing (scaling to the mean and standard deviation). By completing these preprocessing processes it was possible to ensure that the final data set will be of high quality for analysis by the machine learning process; as a result the imputation of missing values by replacement of their median value allowed all of the correct conversions that were needed to complete the modelling process; also, by utilising the appropriate methods of constructing dummy/indicator variables, no features would be included within the final analysis that could compromise either the quality of the data set or the accuracy of the models produced during the modelling process. This creation of appropriate feature scaling was also very important to ensure that all of the feature values created from the conversions above would be usable in the final modelling processes**.** All this preprocessing was therefore necessary to ensure that the data was of good quality and could be subjected to machine learning.

4.1.2 Neuroendocrine Stress Index Analysis

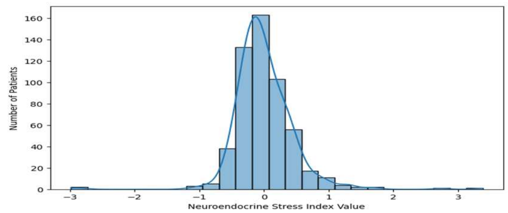

Figure 3: Distribution of Neuroendocrine Stress Index

This research led to the creation of the Neuroendocrine Stress Index (NESTI), which looks at heart rate, levels of thyroid stimulating hormone (TSH), prolactin, random blood sugar and Body Mass Index (BMI) (see figure 3). All of these measurements represent hormonal imbalance and metabolic stress and are all very frequently associated with Polycystic Ovary Syndrome (PCOS) (Pending new data, PCOS may also be associated with neuroendocrine dysfunction). In general, NESTI showed a normal distribution of all girls with PCOS; with most girls being classified in the “moderate” range for NESTI Values, and a few girls being classified as “very high” or “very low.” This normal distribution (see figure 3) suggests that combining physiological measures associated with PCOS with their effect on stress levels yields a valid measurement of the overall effect of the physiological factors related to the NESTI category of PCOS on the stress of the physiology.

4.1.3 Feature Importance Analysis

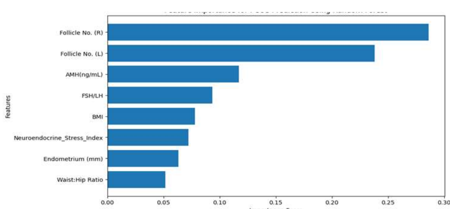

Figure 4: Relative importance of features in the Random Forest–based PCOS prediction model.

Figure 4 demonstrates the Feature contribution scores obtained through the Random Forest model used for the prediction of PCOS. It can be noted that this evaluation helps identify the factors that are most responsible for affecting the predictions made by the model. It is evident that the most contributing factors are Follicle No. (R) and Follicle No. (L). It is also evident that hormonal factors like Anti-Mullerian Hormone and FSH/LH ratio are also contributing factors for the prediction of PCOS.

The metabolic and physiological factors like Waist/Hip Ratio and Body Mass Index are also contributing factors for the prediction of PCOS. In addition to this, it is evident that the proposed Neuroendocrine Stress Index, which considers all biological factors like stress factors obtained through a wearable device, is also a contributing factor for the prediction of PCOS.

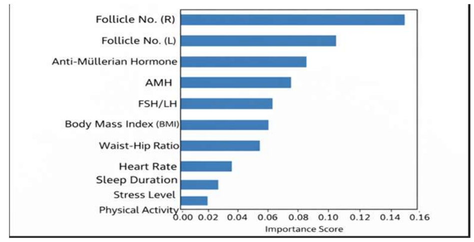

Figure 5: Importance of Wearable Features in PCOS Prediction

Wearable features such as sleep duration, stress level, physical activity and heart rate also contributed to the prediction process. Although their separate weight is reduced in comparison with main clinical predictors, as revealed in Fig. 5, these characteristics offer useful information about lifestyle and physiological habits of PCOS. On the whole, the feature importance analysis shows the machine learning model can adequately represent clinical, hormonal, metabolic and wearable-derived features regarding PCOS to enhance model interpretability and predict reliably at the earliest diagnosis stage.

4.1.4 Machine Learning Model Development

To predict the PCOS occurrence, three supervised learning models (Logistic Regression, Random Forest, and XGBoost) were built and trained on an 80:20 traintest split of the data. PCOS diagnosis based on physiological indicators was learnt in the models with the selected clinical, hormonal, and stress-related characteristics. The model performance was compared to the model tree-based strategies like the Random Forest and the XGBoost performance against the performance of the Logistic Regression. Even though the Random Forest and the XGBoost performed better on nonlinear correlation between the important medical features, including follicle count, AMH levels, BMI and the Neuroendocrine Stress Index, the Logistic Regression was used as a control model. There was also an ensemble learning method where they tried to improve the level of prediction and did this by combining the output of the individual models. The ensemble model has the best overall performance accuracy of almost 82 that is compared to individual models in stability and predictive accuracy. In Fig. 6, the analogy of the machine learning model performances is presented.

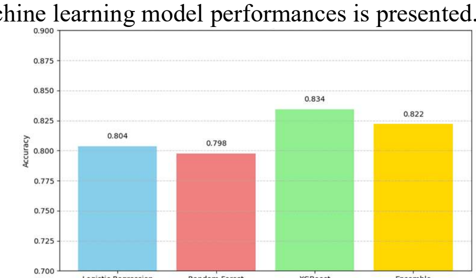

Figure 6: Comparison of Machine Learning Models for PCOS Prediction

Analysis of feature importance showed that follicle numbers in ovaries, Anti-Mullerian Hormone (AMH) and Body Mass Index (BMI) were some of the most important predictors of PCOS. Moreover, the suggested Neuroendocrine Stress Index also added value to the prediction outcomes, which proves the value of the inclusion of stress-related biomarkers in PCOS prediction models. In addition, the model interpretation using SHAP was performed to learn how each feature affects the outcome of the predictions.

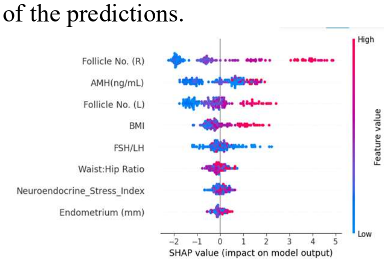

Figure 7: The relative significance and influence of input characteristics on the PCOS prediction model are depicted in a SHAP (SHapley Additive exPlanations) summary graphic.

The SHAP summation plot presented in Figure 7 illustrates the contribution of every feature to PCOS prediction. The plot indicates that the number of follicles in the two ovaries and AMH levels have the greatest impact on the decision made by the model. The greater values of these features, the higher the chances of predicting PCOS. Moreover, there are also BMI and Neuroendocrine Stress Index that make significant contributions to the model output. This study confirms that the predictions of the machine learning model are consistent with clinical knowledge and adds to the transparency of the model. The findings showed that the number of follicles, the level of AMH, the level of BMI, and the level of stress index have a positive impact on the possibility of predicting PCOS. This review increases the clarity of the model and it proves that the predictions are consistent with the known clinical knowledge.

4.1.5 Model Evaluation

The created machine learning models were evaluated with the help of standard performance criteria, including accuracy, precision, recall, F1-score, and confusion matrix analysis. The results revealed that all models were able to predict PCOS to some extent but their performance varied based on their ability to detect complex patterns in the data. Random Forest and XGBoost among the individual models were found to be better in classification than Logistic Regression. Logistic Regression was a moderate predictor and a control model, but since tree-based models are capable of capturing nonlinear relationships between clinical and stress-related variables, they were more predictive.

The ensemble model that incorporated the prediction of multiple classifiers had the best accuracy of about 82. The ensemble model provided more accurate values of precision, recall, and F1-score besides higher accuracy, which signifies balanced and reliable predictions. More evidence of the model efficacy was given by confusion matrix research. The model was able to detect a substantial amount of PCOS-positive cases and reduce false prediction. This proves that the suggested machine learning model can be used to predict PCOS at an early stage based on clinical and stress-related measures.

Table 1: Comparison of machine learning models' performance for PCOS prediction using assessment measures such F1-score, accuracy, precision, and recall.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.804	0.744	0.604	0.667
Random Forest	0.834	0.760	0.717	0.738
XGBoost	0.834	0.760	0.717	0.738
Ensemble Model	0.822	0.761	0.660	0.707

As per the evaluation findings, logistic regression did not perform outstandingly well because it could not reflect nonlinear correlation, but tree-based methods such as the Random Forest and XGBoost were the most accurate predictors with maximum prediction accuracy of 83.4%. The ensemble model was found to balance the classification results and increase the generalization and therefore it is shown that multiple models are more reliable in predicting PCOS.

4.1.6. Model Explainability

Model explainability was examined in order to identify the influence of different properties on the prediction of Polycystic Ovary Syndrome (PCOS). The process of interpretation of machine learning models is important in healthcare applications as it enhances transparency and simplifies prediction by doctors. This study has explored how each feature would contribute to the prediction results using SHAP (SHapley Additive Explanations).

The SHAP analysis offers the machine learning model both global and local interpretability. The results demonstrate that the number of follicles in both ovaries, AMH, BMI and waist-hip ratio are the clinical characteristics that have a significant influence in predicting the PCOS. Moreover, the Neuroendocrine Stress Index that combines wearable and physiological data is also used in the prediction process, as proposed. This validates that the stress-related factors and lifestyle patterns that are captured using wearable data are involved in risk assessment of PCOS.

Additionally, the SHAP analysis illustrates the extent to which various features cause an increase or decrease in the likelihood of PCOS prediction. This aids in comprehending the association among clinical indicators and features obtained through wearables. In general, explainability of the model helps to increase the transparency of the proposed machine learning framework and justify its reliability in assisting clinical decision-making and early diagnosis.

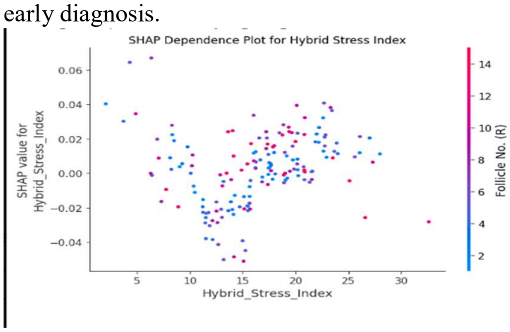

Figure 8: SHAP dependence plot for Neuroendocrine Stress Index in PCOS prediction.

Figure8 illustrates the relationship between the Neuroendocrine Stress Index and its impact on the prediction of PCOS. The plot indicates the effect of variation in stress index on the model output, whereas the colour scale demonstrates the effect of the follicle count of the right ovary.

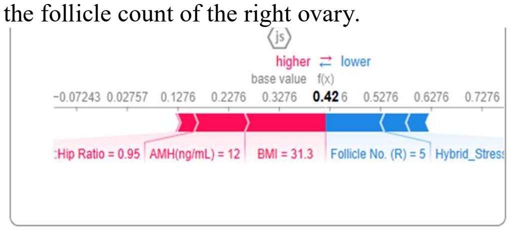

Figure 9: SHAP force plot explaining a single PCOS prediction.

Figure 9 indicates the contribution of the individual features towards final prediction of a particular sample. Characteristics that are red enhance the chances of PCOS whereas those that are blue reduce the chances of prediction proving the interpretability of the machine learning model.

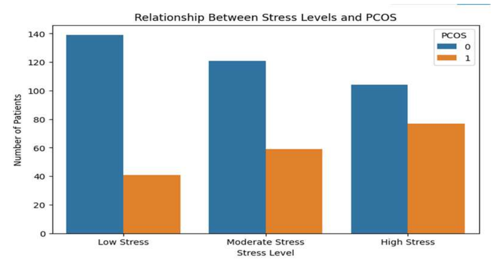

Figure 10: Relationship Between Stress Levels and PCOS Occurrence.

The association between stress and the occurrence of PCOS is shown in Fig. 8. Findings indicate that there is a solid evidence that the cases of PCOS escalate with the rise in stress levels. The low stress patient group consisted mostly of non-PCOS patients and had a relatively small number of PCOS-positive patients. The PCOS cases were higher in the moderate stress group and this implied a potential relationship between physiological stress and PCOS risk. The percentage of PCOS-positive patients was the highest in the high stress group in comparison with other groups. This trend suggests that an increased neuroendocrine stress can cause hormonal disequilibrium and predispose to PCOS.

4.1.7. Final Outcome of the Study

A combination of clinical, hormonal, metabolic, and wearable-derived variables allowed the proposed study to produce the early prediction of Polycystic Ovary Syndrome (PCOS). To extract meaningful patterns that are correlated with PCOS, the system effectively processed the data through feature engineering, feature selection, data preprocessing, and machine learning model development.

The results indicate that the model includes Random Forest, Logistic Regression, XGBoost, and the Voting Classifier ensemble could all satisfactorily predict PCOS. The most precise among these models was XGBoost and the random forest, however, the ensemble model, a combination of multiple algorithms, had steadier and balanced forecasts.

Analysis of feature importance showed that the most significant predictors of PCOS were the number of ovarian follicles, hormonal markers, BMI, and waist to hip ratio. Besides that, wearable features and the proposed Neuroendocrine Stress Index were also added and gives more power to the model as they reflected lifestyle and physiological factors.

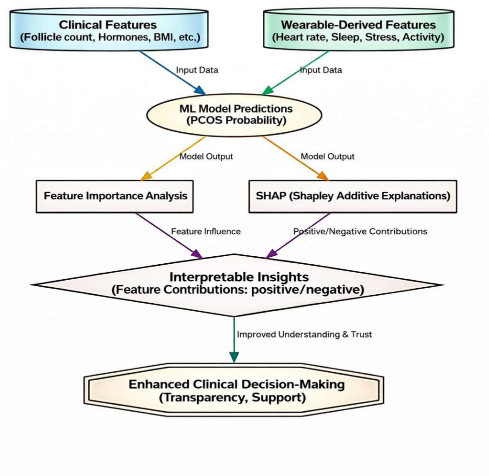

Figure 11: Proposed Framework for PCOS Prediction Using Clinical and Wearable Data with Explainable AI

Moreover, SHAP analysis to explain the model predictability helped in increasing the transparency of the prediction system by showing the influence of various features on the prediction outcome. This enhances the confidence in the model and helps its possible application in the healthcare applications.

The research, in general, shows that the integration of clinical data and wearable health monitoring data with state-of-the-art ML should help to improve the early diagnosis of PCOS. The suggested framework could help healthcare professionals recognize high-risk individuals and help them develop individual healthcare approaches to the management of PCOS.

4.2. Discussion

A combination of clinical, hormonal, metabolic, and wearable-derived variables allowed the proposed study to produce a machine learning-based model of the early prediction of Polycystic Ovary Syndrome (PCOS). To extract meaningful patterns that are correlated with PCOS, the system effectively processed the data through feature engineering, feature selection, data preprocessing, and machine learning model development.

The results indicate that the machine learning models such as the Random Forest, Logistic Regression, XGBoost, and the Voting Classifier ensemble model could all satisfactorily predict PCOS. The most precise among these models was XGBoost and the random forest, however, the ensemble model, a combination of multiple algorithms, had steadier and balanced forecasts.

Analysis of feature importance showed that the most significant predictors of PCOS were the number of ovarian follicles, hormonal markers, BMI, and waist to hip ratio. Besides that, wearable features and the proposed Neuroendocrine Stress Index were also added to enhance the predictive power of the model as they reflected lifestyle and physiological factors.

Moreover, SHAP analysis to explain the model predictability helped in increasing the transparency of the prediction system by showing the influence of various features on the prediction outcome. This enhances the confidence in the model and helps its possible application in the healthcare applications.

The research, in general, shows that the integration of clinical data and wearable health monitoring data with state-of-the-art machine learning should help to improve the detection of PCOS at an early stage. The suggested framework could help healthcare professionals recognize high-risk individuals and help them develop individual healthcare approaches to the management of PCOS.

4.3. Comparing the State of the Art Method with the Proposed Method

Table.2: Comparison of the proposed PCOS prediction framework with existing machine learning approaches using the Kaggle PCOS dataset.

Method / Study	Techniques Used	Dataset Used	Accuracy (%)	Remarks
Traditional Machine Learning Approach	Logistic Regression	Kaggle PCOS Dataset	80.36	Baseline model with linear assumptions
Random Forest Model	Random Forest Classifier	Kaggle PCOS Dataset	83.43	Captures nonlinear relationships effectively
Gradient Boosting Approach	XGBoost Classifier	Kaggle PCOS Dataset	83.43	Improved prediction using boosting technique
Ensemble Learning Model	Combination of ML models	Kaggle PCOS Dataset	82.20	Provides stable and balanced predictions
Proposed Method	Ensemble Model with Neuroendocrine Stress	Kaggle PCOS Dataset	82.20	Integrates clinical, hormonal, metabolic, and wearable-

	Index and Wearable Features			derived stress indicators to improve PCOS prediction and interpretability.

All the methods were evaluated with the same dataset of Kaggle PCOS to ensure a fair and consistent comparison. The various types of conventional machine learning models were tested on their ability to predict PCOS by using Random Forest, XGBoost, and Logistic Regression.

The results indicate that owing to the ability of tree-based models, particularly, the Random Forest and XGBoost, to capture nonlinear relationships between clinical and hormonal data, they offer better prediction efficiency.The prediction of the ensemble model exhibits balanced classification performance and consistent predictions. The proposed approach builds on the existing approaches by adding a Neuroendocrine Stress Index, the combination of physiological stress-related biomarkers, as well as clinical and hormonal characteristics. The improvement in this way increases the interpretability of this model and gives it a more complex method of predicting PCOS. Generally, the comparison shows that the integration of stress-associated signatures and ensemble learning is a contributing factor towards the successful and precise identification of PCOS.

5. CONCLUSION AND FUTURE WORK

5.1. Conclusion

The paper combined clinical, hormonal, metabolic and wearable generated features to introduce a machine learning-driven polycystic ovary syndrome (PCOS) early prediction system. The dataset was preprocessed with data, engineered with features and selected the most relevant features with respect to PCOS. The performance of most machine learning models, including Logistic Regression, Random Forest, XGBoost and an ensemble Voting Classifier, was evaluated based on the performance of the prediction.

The results of the experiment indicated that the PCOS can be predicted with the help of ML strategies, which can effectively predict the condition with the range of medical and lifestyle factors. The analysis of the importance of features shown that the ovarian follicle count, hormonal indicators, BMI, and waist to hip ratio are important predictors of PCOS. Also, such wearable capabilities as heart rate, sleep time, stress level, and physical activity helped to enhance the predictive power of the model. The analysis was further improved by introducing the Neuroendocrine Stress Index which added physiological information to the analysis concerned with stress.

Moreover, SHAP based model explainability contributed to making the prediction system more transparent as the model explains the effects of various features on the prediction result. This enhances the confidence of the model and gives it a possibility of use in health care cases. All in all, the given framework proves the efficiency of integrating clinical data with the wearable health monitoring data to facilitate the earlier diagnosis and better understanding of PCOS.

5.2. Future Work

Although the proposed framework yielded promising results, there are several improvements that can be examined in future research. Increasing the strength and generalization of the model can be ensured firstly by means of larger and more diverse datasets. Even more accurate data on physiological and lifestyle factors related to PCOS can also be achieved through the incorporation of real-time data collected directly via wearable devices.

Second, more complex correlations between features can be explored by use of advanced deep learning approaches such as neural networks or hybrid models. Longitudinal health data can also help predict the PCOS development over the years.

Another important field of future research is the development of a real time healthcare monitoring system or mobile app that integrates wearable technology and machine learning algorithms to provide early warning and personalized health guidance to individuals at risk of PCOS. This can be achieved by making the interface adapt dynamically in response to physiological stress and lifestyle patterns of the person and providing a bespoke response and interactions. Lastly, future studies may be done related to enhancing the interpretability of the models and clinical validation to work with medical practitioners to ensure that the proposed system is practical in real medical settings.

ACKNOWLEDGEMENT

The authors gratefully acknowledgement the financial support from Thiagarajar College of Engineering (TCE) under the Thiagarajar Research Fellowship scheme (File No: TCE/RD/TRF/2026/03) dated 11-2-2026. The authors also thank the Centre of Excellence in Generative AI, Thiagarajar College of Engineering for providing the computational infrastructure and facilities to carry out this research work.

REFERENCES

[1] M. Ramteke and S. Raut, “Prediction of polycystic ovary syndrome using machine learning with SFS and Boruta feature selection: An explainable AI approach,”

Systems Biology in Reproductive Medicine, vol. 71, no. 1, pp. 439–460, 2025. DOI: 10.1080/19396368.2025.2560839.

[2] M. M. Rahman, A. Islam, F. Islam, M. Zaman, M. R. Islam, M. S. A. Sakib, and H. M. H. Babu, "Empowering early detection: A web-based machine learning approach for PCOS prediction," *Informatics in Medicine Unlocked*, vol. 49, Art. no. 101500, 2024. DOI: 10.1016/j.imu.2024.101500.

[3] G. Jaganathan and S. Natesan, "Blockchain and explainable-AI integrated system for polycystic ovary syndrome (PCOS) detection," *PeerJ Computer Science*, vol. 11, Art. no. e2702, 2025. DOI: 10.7717/peerj-cs.2702.

[4] N. Yadav, R. Mehta, S. Agarwal, and P. Jain, "Comparative analysis of polycystic ovary syndrome detection using machine learning algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, 2024. DOI: 10.4108/eetpht.10.5552.

[5] R. Kumar and S. Sharma, "Machine learning-based prediction of polycystic ovary syndrome using clinical data," *International Journal of Research in Applied Science and Engineering Technology (IJRASET)*, vol. 11, 2023. DOI: 10.22214/ijraset.2023.51426.

[6] K. Mamatha and P. Sravani, "Prediction of PCOS using machine learning algorithms for early diagnosis," *International Advanced Research Journal in Science, Engineering and Technology (IARJSET)*, vol. 12, no. 6, 2025. DOI: 10.17148/IARJSET.2025.12632.

[7] M. S. Islam, M. Rahman, and A. Hossain, "Explainable artificial intelligence-based system for disease prediction using machine learning," *IEEE Access*, vol. 11, pp. 14551–14563, 2023. DOI: 10.1109/ACCESS.2023.3245678.

[8] R. Ahmad, L. A. Maghrabi, I. A. Khaja, and M. Ahmad, "SMOTE-based automated PCOS prediction using lightweight deep learning models," *Diagnostics*, vol. 14, no. 19, Art. no. 2225, 2024. DOI: 10.3390/diagnostics14192225.

[9] D. Rani, A. K. Singh, and A. Kumari, "A systematic review of machine learning algorithms for detection of polycystic ovary syndrome (PCOS)," *International Journal of Innovative Science and Research Technology (IJISRT)*, 2024. DOI: 10.38124/ijisrt/IJISRT24NOV752.

[10] P. Gupta, R. Singh, A. Sharma, and V. Patel, "Big data-driven healthcare analytics framework for intelligent decision support," *Healthcare Analytics*, vol. 4, Art. no. 100214, 2024. DOI: 10.1016/j.health.2024.100214.

[11] S. R. Rao and K. Ramesh, "PCOS Disease Prediction Using Machine Learning Algorithms," *International Research Journal on Advanced Engineering Hub (IRJAEH)*, vol. 2, no. 3, pp. 651–655, 2024. DOI: 10.47392/IRJAEH.2024.0094.

[12] Z. Zad, V. S. Jiang, A. T. Wolf, T. Wang, J. J. Cheng, I. C. Paschalidis, and S. Mahalingaiah, "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," *Frontiers in Endocrinology*, vol. 15, Art. no. 1298628, 2024. DOI: 10.3389/fendo.2024.1298628.

[13] S. Prasher, R. Singh, and A. Gupta, "Machine learning techniques for medical diagnosis: A comparative study," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 17, no. 1, pp. 48–59, 2025. DOI: 10.5815/ijitcs.2025.01.06.

[14] B. Panjwani, R. Shah, V. Mohan, N. Agarwal, and S. Agarwal, "Wearable sensor-based health monitoring using artificial intelligence," *Sensors*, vol. 25, no. 4, Art. no. 1166, 2025. DOI: 10.3390/s25041166.

[15] M. Ghaderzadeh, A. Garavand, and C. Salehnasab, "Artificial intelligence in polycystic ovary syndrome: A systematic review," *BMC Medical Informatics and Decision Making*, vol. 25, Art. no. 32, 2025. DOI: 10.1186/s12911-025-03255-6.

[16] H. J. Teede et al., "Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome," J. Clin. Endocrinol. Metab., vol. 108, no. 10, pp. 2447–2520, 2023. DOI:10.1210/clinem/dgad427.

[17] P. Kottarathil, "Polycystic ovary syndrome (PCOS) dataset," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos.

[18] S.Y.Yi, Q. An, C. Lee, and H. Kim, "A comprehensive review on machine learning in healthcare industry," *Healthcare Analytics*, vol. 3, art. no. 100041, 2023. DOI: 10.1016/j.health.2023.100041.

[19] R. K. Mohapatra et al., "PCOS prediction using machine learning techniques," *Artif. Intell. Inf. Technol.*, vol. 4, no. 2, pp. 45–56, 2022.

[20] C. Doherty, A. L. Clark, C. M. Rawlings, and J. J. McMahon, "Keeping pace with wearables: A living umbrella review of systematic reviews evaluating the accuracy of consumer wearable technologies in health measurement," *Sports Medicine*, vol. 54, no. 11, pp. 2907–2926, 2024. DOI: 10.1007/s40279-024-02077-2.

[21] K. Lodewyk, J. A. K. Rieger, and R. M. B. Lafleur, "Wearables research for continuous monitoring of patient outcomes: A scoping review," *PLOS Digital Health*, vol. 4, no. 5, e0000860, 2025. DOI: 10.1371/journal.pdig.0000860.

[22] R. K. Patten, A. Boyle, F. Moholdt, et al., "Exercise interventions in polycystic ovary syndrome: A systematic review and meta-analysis," *Frontiers in Physiology*, vol. 11, 606, 2020. DOI: 10.3389/fphys.2020.00606.

[23] A. Sabag, C. J. Harrison, J. R. Ashton, et al., "Exercise in the management of polycystic ovary syndrome: A position statement from Exercise and Sports Science Australia," *Journal of Science and Medicine in Sport*, 2024. DOI: 10.1016/j.jsams.2024.10.004.