Enhancing PCOS Detection: A Hybrid Approach Utilizing Blood Profiles and Ultrasound Imaging

Shatakshi Shree[1*] and Dr. Sandosh S[2†]

[1*]CSE with specialisation in AI and ML, Vellore Institute of Technology, Chennai, 600127, Tamil Nadu, India.
[2]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, 600127, Tamil Nadu, India.

*Corresponding author(s). E-mail(s): shatakshi.shree2020@vitstudent.ac.in;
Contributing authors: sandosh.s@vit.ac.in;
†These authors contributed equally to this work.

Abstract

Polycystic Ovary Syndrome (PCOS) is a common condition affecting women's ovaries, leading to hormonal imbalances during their reproductive years. This imbalance results in various issues such as irregular menstrual cycles, acne, hair loss, thinning of hair, cramps, insulin resistance, metabolic syndrome, infertility, and mood swings. Despite its widespread occurrence, many cases go undetected, and PCOS is increasingly being recognized as a psychosomatic disorder. This research is a comprehensive investigation into the diagnosis of PCOS, employing an integrated approach utilizing both, blood parameters and ultrasound images. For blood-based diagnostics, machine learning algorithms like random forest and hyperparameter tuning using RandomizedSearchCV and GridSearchCV, Decision Tree with cost complexity pruning, Logistic regression, KNN, Naïve Bayes, Neural network, LDA, QDA, Nearest centroid classifier, Gaussian process classifier, Full grown tree, Voting classifier (with logistic regression, Decision Tree, and SVM), Bagging classifier, Extratrees classifier, Adaboost classifier, XGBoost have been used which revealed patterns emphasising the significance of hormonal imbalances and metabolic parameters. Concurrently, ultrasonographic evaluations provided insights into ovarian morphology, emphasizing the presence of follicles and other structural abnormalities. For differentiating between the images of infected and non-infected patients CNN with softmax activation in the output layer and ResNet50 have been used. This offers a deeper insight into the multifaceted nature of PCOS. There have been reports where PCOS is not detected in

the ultrasound images but shows its presence in the patient's body by deranging the hormones. Thus, we integrated both the findings to find a model that can detect PCOS with more sensitivity and specificity.

Keywords: PCOS, random forest, CNN, blood markers, ultrasound imaging, hormonal imbalances, ovarian morphology, decision tree, Adaboost, XGboost, KNN, LDA, QDA, ResNet50, Voting classifier

1 Introduction

Polycystic Ovary Syndrome (PCOS aka PCOD) is a multifaceted clinical problem, with symptoms often including irregular menstrual cycles, hyperandrogenism, and polycystic ovaries. It also poses a diagnostic challenge for healthcare providers today as it is a non-confirmatory disease or more like a psychosomatic disorder. Early and accurate detection is a need for prevention of associated complications such as infertility, cardiovascular diseases, and metabolic disturbances. Traditional diagnostic methods often rely either on blood profile analyses or ultrasound to identify hormonal imbalances associated with PCOS. While informative, these methods may lack the necessary specificity and sensitivity, leading to potential misdiagnoses and delayed interventions. Additionally, the sole reliance on blood profiles overlooks the structural aspects of ovarian morphology that can be crucial for accurate PCOS diagnosis. Concurrently, ultrasound imaging serves as a valuable tool for visualizing ovarian morphology, including the presence of cysts. However, the exclusive reliance on ultrasound imaging may miss nuanced variations in ovarian characteristics and may not provide a comprehensive understanding. PCOS is now seen not just as a disease but a lifestyle disorder that can be controlled with a healthy diet and active life. In the current medical landscape, there is a growing recognition of the need for more nuanced and integrated diagnostic approaches. Advances in machine learning, image processing, and artificial intelligence present an opportune avenue for developing hybrid models that combine the strengths of blood profile analyses and ultrasound imaging. It being incurable is a problem but so is life. But, it can be maintained and life can be sustained by taking medications and maintaining a good lifestyle While various diagnostic criteria have been established, there exists a pressing need for advanced and integrated methodologies that enhance the precision and efficiency of PCOS detection. This research aims to address this by proposing a novel hybrid approach that handles both blood profiles and ultrasound imaging for comprehensive PCOS diagnosis.

The first part of the investigation focuses on blood profile analysis which have clinical and metabolic features, delving into the intricate significant features that bear the fingerprints of PCOS. Recognizing the limitations of existing methods, the research aims to contribute to the field by employing a diverse and representative dataset, implementing machine learning algorithms like Random Forest, Logistic regression, Decision Trees, Voting classifier, Bagging classifier, Extra trees, Adaboost classifier, KNN, MLP, Gaussian Naïve Bayes model, Linear Discriminant Analysis, Nearest Centroid model and Quadratic Discriminant Analysis. By synergistically combining the

strengths of multiple models, the approach aspires to achieve heightened accuracy in the identification of PCOS through blood profiles.

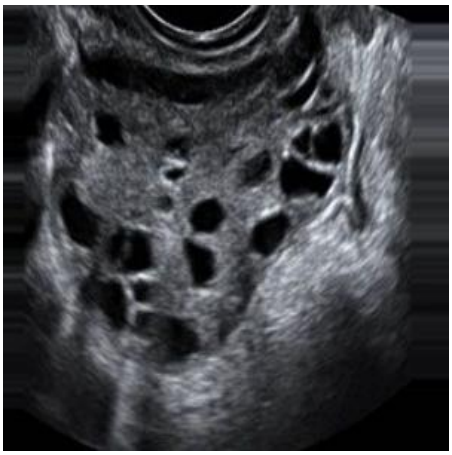

Fig. 1 USG of a patient with PCOS

Fig. 2 USG of a healthy woman

In the second dimension of the study, attention is given to medical imaging, specifically ultrasound (USG). Acknowledging the crucial role ultrasound plays in visualizing ovarian cloggy and identifying characteristic cysts and follicles visible in the ultrasound. For a sophisticated analysis, two models have been used, the first being CNN compiled with Adam optimizer and binary cross-entropy loss and the second being transfer learning with ResNet50 architecture. After consulting experienced medical practitioners, we got to know that PCOS is non-confirmatory. The goal standard of detecting PCOS is still unknown as there is a lack of well-defined diagnostic criteria. The practitioners' way of detecting PCOS is by looking for the most important clinical features first which are – menstrual disturbances, hirsutism, acne, and anovulatory infertility. If someone possesses the common signs, they go for an ultrasound (USG) of the pelvis. If there exist small sacs called follicles, then it gets confirmed over there itself. But, if the USG report is normal then they do other tests like checking LH-FSH levels, fasting glucose/insulin (should be less than 4.5), testosterone, and free androgen levels. If the patient has PCOS on the ultrasound but no other signs like too much male hormones or period irregularities, then she doesn't have PCOS and thus, doesn't need any further tests. According to the Rotterdam criteria many patients can be diagnosed based on their history and physical features, like having irregular periods (menstruation) and signs of too much male hormone (like acne, extra hair, or hair loss). The test 'AMH (ng/mL)' is not commonly used to check for PCOS right now. Most experts use the Rotterdam criteria for its diagnosis. Two out of three of the following criteria are required to make the diagnosis:

- Oligo- and/or anovulation (irregular ovulation and/or absence on ovulation)
- Clinical and/or biochemical signs of hyperandrogenism

- Polycystic ovaries (by ultrasound)

As the blood reports contain the clinical features or the symptoms and also the biochemical tests, the ultrasound image diagnosis, and the blood reports are to be given equal importance for accurate results. Thus, they have 50-50 significance. If even then things are not clear then the patient must go for additional tests of testosterone and androgen levels, in cases like ultrasound images support PCOS but the blood reports analysis opposes it or vice-versa.

This comprehensive investigation enhances the diagnostic accuracy of PCOS and also contributes to the evolving landscape of precision medicine. By amalgamating the insights derived from blood profiles and the visual nuances captured through ultrasound imaging, the hybrid approach seeks to provide a holistic and nuanced understanding of PCOS, fostering improved patient outcomes and more tailored therapeutic strategies. This study seeks to operate at the intersection of traditional diagnostics and cutting-edge technology, aiming to bridge the existing gaps in PCOS detection methodologies. By employing a hybrid approach, the aim is to harness the complementary information embedded in blood profiles and ultrasound images, creating a more robust and comprehensive diagnostic framework for PCOS. This study holds importance as it has the potential to change how PCOS is diagnosed by overcoming the drawbacks of existing methods.

1.1 Objective

he objective of the proposed work is:

1. To find a model that achieves decent accuracy for the prediction of PCOS for the blood reports which has both metabolic and clinical features.
2. To utilize the USG reports of patients for the detection of PCOS.
3. Integrating both modules to get a reliable result that features a combination of both studies.

1.2 Organization

he paper has been divided into three sections which are as follows: Section 2 provides an overview of the existing works in this field. Section 3 provides a comprehensive explanation of the framework and algorithms proposed for Blood Profile analysis and USG analysis. The analysis of the performance of the proposed work is demonstrated in Section 4. Lastly, Section 5 summarises the research findings and outlines potential future directions for this work.

2 Literature Survey

The study [1] analyses 541 cases, with 364 classified as normal and non-PCOS and 177 as PCOS. After PCA, eight potential features are identified as significant for discriminating between PCOS and non-PCOS. Random Forest Classifier emerges as the most accurate model, achieving 89% accuracy. The results underscore the limitation of

relying solely on biochemical profiles or ultrasound results for PCOS diagnosis. Anti-Müllerian Hormone (AMH) is identified as a promising feature for PCOS detection. While the accuracy of this study (89%) is slightly lower than some previous research, the potential for improvement through classifier weight parameter optimization is acknowledged.

The research paper [2] suggests the usage of computer-based methods to diagnose PCOS/PCOD in women. The authors have used a dataset from Kaggle containing data of 541 women with 43 different attributes. It also uses a method called uni-variate feature selection to figure out which features are most useful for predicting PCOS. In their study they found that the ratio of FSH and LH are the most important factor and thus concentrated their study on it. They try out different ML algorithms, such as gradient boosting, random forest, logistic regression, and a combination of random forest and logistic regression (RFLR). The RFLR method turns out to be the best, giving a testing accuracy of 91.01% and a recall value of 90%, using a technique called 40-fold cross-validation on the top 10 features. The paper stresses the importance of early screening for PCOS and highlights how carefully choosing which features to look at can make the diagnosis more accurate and faster. They compare their results with other research and showcase that their RFLR method performs well even with only 10 features, making it quicker to process saving computational resources.

The research paper [3] tackles the significant concern of Polycystic Ovarian Syndrome (PCOS), an endocrine disorder impacting 5-10% of females in their teenage. With the aim of improving early detection and treatment, the study employs RapidMiner to evaluate models and features selection on the Kaggle dataset. Random Forest emerges with the highest accuracy (93.12%, RapidMiner) on the complete dataset. KNN and SVM perform similarly (90.83%, RapidMiner) with 10 selected features. Feature selection analysis indicates that 10 features can yield competitive results compared to the complete set, while 24 features show significant differences. The study compares the performance of Python and RapidMiner, revealing that RapidMiner outperforms in accuracy, precision, and recall. However, the paper acknowledges the dependency on the dataset's nature and the techniques used. Overall, the research emphasizes the potential of machine learning in PCOS prediction and underscores the importance of tool and algorithm selection based on dataset characteristics.

The research paper [4] addresses this critical issue by introducing an automated PCOS detection method utilizing clinical and metabolic parameters as early markers for the disease. The proposed algorithm employs a feature vector based on these parameters, with statistically significant features selected using a two-sample t-test. Bayesian and Logistic Regression classifiers are utilized for classification, demonstrating that the Bayesian classifier outperforms logistic regression with an accuracy of 93.93% compared to 91.04%. The study emphasizes the potential of the automated system as an assisting tool for doctors, saving time in patient examinations and aiding in early PCOS risk diagnosis. Additionally, the research contributes by highlighting the importance of clinical features such as FSH, LH, BMI, and cycle length in discriminating between normal and PCOS groups. Overall, the study underscores the

effectiveness of the Bayesian classifier for early PCOS screening, suggesting its superiority over logistic regression in this context and advocating for further improvements in accuracy through the exploration of alternative classifiers.

The study [5] emphasizes on the need for enhanced predictive models to identify Polycystic Ovary Syndrome (PCOS) at an early stage, especially among a wider outpatient population, with a specific focus on individuals who have signs of the problem but have not received a diagnosis yet. Utilizing machine learning algorithms on electronic health records (EHR) from a SafetyNet hospital covering 30,601 women, the research developed predictive models using logistic regression, support vector machine, gradient boosted trees, and random forests. The models integrated hormone values and other clinical parameters to predict PCOS outcomes. The predictive accuracy of these models, assessed by the area under the curve (AUC), demonstrated promising results ranging from 80% to 85% in out-of-sample test sets. Important factors that indicated a high likelihood were hormone levels and obesity, while factors indicating a lower likelihood were the number of pregnancies and a positive bHCG. The study highlights tools integrated into EHR to facilitate early PCOS detection, counseling, and interventions to mitigate long-term health consequences. However, the need for model validation in diverse hospital-based populations is emphasized.

In the research [6] Polycystic Ovary Syndrome (PCOS) poses a significant health challenge for women in their childbearing years due to hormonal imbalances, leading to various symptoms such as irregular menstrual cycles, excessive weight gain, facial hair growth, acne, and infertility in severe cases. The current diagnostic and treatment approaches lack effectiveness in early-stage detection and prediction. To address this, our proposed system utilizes five machine learning classifiers. They employed the CHI SQUARE method to identify the top 30 features out of 41 in the dataset, forming the feature vector. Comparative analysis of classifier results revealed Random Forest as the most accurate and reliable. The dataset, owned by Prasoon Kottarathil and available on KAGGLE, served as the basis for training and testing our system, offering a promising avenue for enhancing PCOS diagnosis and prediction.

In the study [7] Polycystic ovary syndrome (PCOS) is a significant health concern affecting women's fertility and overall well-being, emphasizing the importance of early diagnosis for effective treatment. Recent advancements in medical diagnosis leverage machine learning methods, which have shown promising results. Additionally, feature selection techniques, focusing on extracting the most relevant subset of features, play a crucial role in enhancing computational efficiency and classifier performance. It examines how both traditional and combined computer-based models can be used on the kaggle PCOS dataset. They have tested various combined models using the dataset with all its features and with selected smaller sets of features chosen through different methods. The results show that how choosing the right feature can significantly impact the performance of the models and thus improving it. Surprisingly, the Ensemble Random Forest model, which used a smaller set of features picked through one of the methods, performed better than the others, achieving an Accuracy of 98.89% and Sensitivity of 100%.

They [8] utilize machine learning (ML) techniques to identify key clinical and laboratory variables associated with Polycystic Ovary Syndrome (PCOS) diagnosis and to

stratify patients into distinct phenotypic clusters. The dataset consists of 72 patients with PCOS and 73 healthy women. To predict and group them, the BorutaShap method and Random Forest algorithm are used. Out of the 58 factors studied, the algorithm highlights specific ones like lipid accumulation product (LAP), abdominal circumference, thrombin activatable fibrinolysis inhibitor (TAFI) levels, body mass index (BMI), C-reactive protein (CRP), high-density lipoprotein cholesterol (HDL-c), follicle-stimulating hormone (FSH), insulin levels, HOMA-IR value, age, prolactin, 17-OH progesterone, and triglycerides levels, as well as a family history of diabetes mellitus in a first-degree relative, concerning the diagnosis of PCOS. When these factors are used together, they achieve an accuracy of 86% and an area under the ROC curve of 97%. This algorithmic approach not only identifies crucial variables for PCOS but also classifies patients into phenotypically different clusters, potentially guiding more personalized and effective treatment strategies for PCOS.

This study [9] addresses the complex disorder by leveraging Artificial Intelligence (AI) through heterogeneous ML and DL classifiers for the prediction of PCOS in fertile patients. Using an open-source dataset from Kerala, India, consisting of 541 patients, the study employs a multi-stack of ML models, achieving outstanding performance with an accuracy, precision, recall, and F1-score of 98%, 97%, 98%, and 98%, respectively. The integration of Explainable AI (XAI) techniques, including SHAP (SHapley Additive Values), LIME (Local Interpretable Model Explainer), ELI5, Qlattice, and feature importance with Random Forest, enhances the transparency and interpretability of the model.

This study [10] focuses on the important issue of Polycystic Ovary Syndrome (PCOS/PCOD) in women of reproductive age (10-12 years) and aims to predict PCOS using advanced ML techniques. PCOS, characterized by hormonal imbalance and ovarian dysfunction, can result in many problems in later stages of life. The research uses a dataset containing clinical and physical parameters of women and introduces a new approach for selecting relevant features, known as the optimized chi-squared (CS-PCOS) mechanism. To compare different models, 10 hyper-parameterized machine learning models are employed, and the gaussian naive bayes (GNB) model stands out, achieving 100% accuracy, precision, recall, and F1-scores with minimal time computations of 0.002 seconds. The k-fold cross-validation of GNB also attains a 100% accuracy score, showcasing its strong and consistent performance. The proposed model highlights features such as prolactin, blood pressure, thyroid stimulating hormone, relative risk, and pregnancy, contributing to the early and accurate prediction of PCOS. The aim of this research is to assist the medical community in reducing miscarriage rates and providing timely interventions for women through the early detection of PCOS.

This research [11] addresses this significant issue as a complex hormonal and reproductive ailment affecting approximately 10 million women globally. In India, it is estimated to afflict one in every five women. Early diagnosis is crucial as PCOS can lead to severe complications. The study aims to identify diseases that may serve as indicators of PCOS, contributing to its early identification. By amalgamating datasets related to identified diseases, performing feature selection, and applying supervised and unsupervised learning algorithms, the research identifies key features for PCOS prediction. The findings emphasize the relevance of obesity, heart disease, high blood

pressure, and diabetes in early PCOS detection. The study's contributions include creating a new dataset, applying various learning algorithms, and providing insights for healthcare professionals to facilitate early PCOS detection with minimal features. The article concludes with a discussion of results, future work, and a list of references.

This research [12] addresses the challenge of early and accurate diagnosis of Polycystic Ovarian Syndrome (PCOS), a prevalent gynecological disorder affecting women globally. Due to the economic impact of PCOS and its associated health issues, there is a need for effective diagnostic tools, especially for adolescent women. The study evaluates the performance of machine learning (ML) algorithms in screening PCOS using 23 non-invasive parameters. The dataset comprises clinical data from 540 patients in Kerala, India, with 378 used for training and 162 for testing. The research aims to identify suitable ML algorithms for PCOS screening without invasive tests. The study demonstrates the promising potential of ML models as effective screening tools for PCOS, showcasing notable sensitivity and ROC performance.

In the study [13] it can be concluded that Polycystic Ovary Syndrome (PCOS) poses a significant threat to women's health, often causing lifelong damage. Many women suffer without realizing they are affected or if the condition is not caught in its early stages. PCOS is a treatable cause of infertility and can impact a woman's health in various ways, leading to conditions like metabolic syndrome, sleep apnea, depression, and even endometrial cancer. However, with early detection and careful supervision, these consequences can be avoided. This research employs various machine learning methods to establish an efficient decision tree, identifying the best-performing model. Once a woman learns from a doctor that she has PCOS, she can subsequently monitor key noticeable changes in her body and hormonal levels using the decision tree. This empowers her to make informed decisions about when to seek medical attention, ensuring better control over her health.

This study [14] aims to use ML algorithms to classify and predict this syndrome based on radial pulse wave parameters. The goal is to provide evidence supporting the objective measurement of pulse diagnosis in traditional Chinese medicine (TCM). In a case-control study involving 459 subjects, including a PCOS group and a healthy (non-PCOS) group, pulse wave parameters were measured and analyzed. Seven supervised ML classification models were applied. Parameters that significantly differed between the PCOS and healthy groups were taken as the input features, and the models underwent stratified k-fold cross-validation training. The PCOS group comprised 316 subjects, while the healthy group had 143 subjects. Certain pulse wave parameters showed significant differences between the two groups. Among the ML models evaluated, both Voting and LSTM, which have ensemble learning capabilities gave good results. They achieved the highest results across all evaluation metrics, with a testing accuracy of 72.174% and an F1 score of 0.818. The respective AUC values were 0.715 for Voting and 0.722 for LSTM. The technique also has the potential to revive the development of personalized PCOS risk assessment using mobile detection technology and also offers physicians an understanding of objective pulse diagnosis in TCM.

[15] The hormonal disorder PCOS affects a significant population of fertile women globally, with documented positive cases ranging from 2.2-26% worldwide. The existing diagnostic methods are often time-consuming, labor-intensive, and prone to errors.

This paper introduces an innovative automated diagnostic system designed for efficient PCOS prognosis using machine learning on clinical data by incorporating feature selection through Particle Swarm Optimization (PSO) followed by a modified stacked generalization ensemble learning model. The proposed system showcases an impressive accuracy of 90.74%, surpassing the performance of previously suggested diagnostic systems.

3 Proposed Work

This section provides a detailed description of the proposed system architecture designed for the hybrid model made for enhancing PCOS detection. It consists of two primary modules – the Blood profile analysis module and the ultrasound imaging module. These modules operate in tandem, leveraging machine learning algorithms and image processing techniques, respectively. The culmination of these modules is the Hybrid Approach Integration, forming a robust diagnostic framework for PCOS.

3.1 Blood Profile Analysis

he proposed system aims to enhance the detection of Polycystic Ovary Syndrome (PCOS) by integrating data from the datasets "PCOS_infertility.csv" and "PCOS_data_without_infertility.xlsx," the system ensures a comprehensive analysis of both fertile and infertile patient records. Through a series of preprocessing steps, such as data merging, handling data types, and addressing missing values, the system ensures the quality and integrity of the dataset.

3.1.1 Data collection and Preprocessing

he dataset is taken from Kaggle and has 44 physical and clinical parameters to detect PCOS and infertility related problems. The data was collected from 10 different hospitals across Kerala, India. The dataset has two sub-data folders named 'PCOS_data_without_infertility.xlsx' and 'PCOS_infertility.csv'. The initial step involved importing important packages and loading the two datasets mentioned above. The two datasets are then merged based on the common feature 'Patient File No.' to consolidate information on patients with and without infertility. Subsequently, repeated features such as 'Unnamed: 44', 'Sl. No_y', 'PCOS (Y/N)_y', 'I beta-HCG(mIU/mL)_y', 'II beta-HCG(mIU/mL)_y', and 'AMH(ng/mL)_y' were removed after merging.

3.1.2 Data Cleaning

he columns 'AMH' and 'II beta-HCG' had dtype objects so encoded the categorical variables. In the database the type objects were numeric values saved as strings thus, it was converted into a numeric value. Encountered an 'a' value in 'AMH(ng/mL)' so we converted all the values to numeric and the value that cannot be converted to numeric would be replaced with NaN. There were missing values in the columns - 'Marriage status' & 'Fast food'. Filled NA values with the median of that feature because there was not much difference in the 'mean' and 'median' of the 'Marraige Status' column

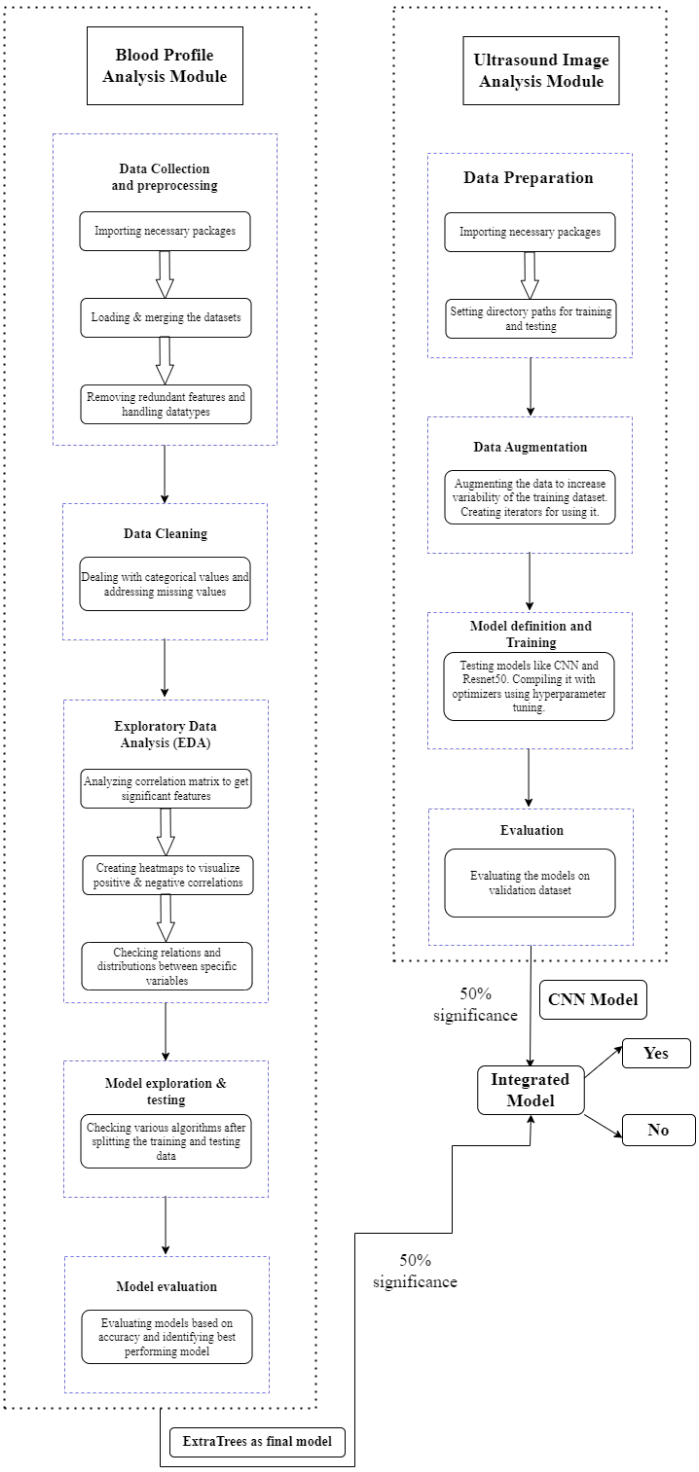

Fig. 3 Model Architecture of the Proposed Work

and thus for others as well. As the number of '1s' in the 'Fast Food' column was we assumed that the person with missing/NaN value eats fast food or we can just put the median value as well so we replaced those values with median. Extra spaces in column names were cleared, and irrelevant columns ('S.No.' and 'Patient File No.') were dropped.

3.1.3 Exploratory Data Analysis (EDA)

t was commenced with the generation of a correlation matrix to discern relationships between variables, particularly focusing on their impact on the target column 'PCOS (Y/N).' Patterns of the menstrual cycle, BMI, irregularity in menstruation, no. of follicles, etc was studied in this. It was noticed that the duration of the menstrual phase remains relatively stable across various age groups in typical instances. However, for individuals with Polycystic Ovary Syndrome (PCOS), the length of this phase tends to extend as age increases. The Body Mass Index (BMI) remains stable in regular cases, while it tends to rise with age in cases of Polycystic Ovary Syndrome (PCOS). In normal instances, the menstrual cycle becomes more consistent over time, but in PCOS cases, irregularities tend to increase as age advances. Additionally, the distribution of follicles between the left and right ovaries differs in women with PCOS compared to those considered "Normal." Women diagnosed with PCOS commonly exhibit an elevated follicle count, as expected, and this count is often unevenly distributed between the ovaries with the follicles bigger than average in size. Later a heatmap was constructed to visualise positive and negative correlations. A distribution analysis of all variables was conducted to gain insights into their characteristics. Specifically, the relationship between 'Follicle (right)' and 'Follicle (left),' identified as having the highest correlation with the target variable, was studied.

3.1.4 Model Evaluation

he dataset was partitioned into training and testing sets (80% and 20%, respectively). Various machine learning algorithms, including random forest, randomized search CV, grid search CV, cost complexity pruning, logistic regression, KNN, Naive Bayes, neural network, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Nearest Centroid Classifier, Gaussian Process Classifier, Decision tree, Full Grown tree, voting classifier (SVM, Logistic regression, decision tree), bagging classifier, Adaboost classifier, and XGboost Classifier were employed. Certain algorithms gave the best accuracies among the others. They were – Logistic regression which is a linear model for binary classification that predicts the probability of the instance belonging to a particular class using the logistic function to map any real-valued number into the range of [0, 1]; Voting classifier which combines the predictions of multiple base estimators and predicts the class with the majority vote; Bagging Classifier which involves aggregation of predictions from multiple base models; Random Forest Classifier which creates multiple decision trees during the training and then gives mode of the classes as output; LDA which is a classification algorithm assuming a common covariance matrix for all classes. Model performance was evaluated through an assessment of their accuracy. The highest accuracy was given by ExtraTreesClassifier which is an ensemble learning method belonging to the family of decision tree-based algorithms.

It features the idea of randomness a step further as it selects a feature and threshold of split randomly, thus, making it more robust to overfitting and increasing the diversity among the trees. As the splitting is done without searching for the best split, it is computationally less extensive compared to the traditional decision tree and random forest. The optimal model was identified using the following hyperparameters: n_estimators=600, max_depth=20, n_jobs=-1, and random_state=0. The model was fitted on the entire dataset (X, y), yielding an accuracy of 99.81%.

3.2 Ultrasound Imaging Analysis

n this section, we delineate the systematic approach undertaken to detect Polycystic Ovary Syndrome (PCOS) in ultrasound images. The process involves several key steps encompassing data collection, preprocessing, model architecture, and validation. Each step is meticulously designed to ensure the reliability and generalization of the Convolutional Neural Network (CNN) and ResNet50 model.

3.2.1 CNN Model

t starts with the data collection and preprocessing which involves steps like importing essential libraries like NumPy, TensorFlow, ResNet50 model, and Matplotlib which provided a robust foundation for numerical, DL implementation and visualisation. To bring reproducibility, a random seed was set to 12 within the TensorFlow environment. The parameters like batch size and image dimensions are then set to 32 and 224x224 respectively. The dataset is taken from Kaggle and has a size of 132MB. The data folder has 'train' and 'test' subfolders containing 2 categories of data 'infected' – images of ovaries having PCOS and 'notinfected' – images of healthy ovaries making a total of 3856 images. After labelling the dataset in the training directory, it is split into training and validation sets for model training. To gain a better understanding of the dataset we visualised a sample of images. Then to avoid data scarcity, we augmented the data. This step is done to enhance the model generalisation. The employed data augmentation techniques were rescaling, rotation, and flipping thus increasing the diversity of training dataset. This step remains common for both the models, i.e. CNN and ResNet50.

Data generators were created for both training and validation datasets as they provide real-time data augmentation during model training. CNN architecture was implemented using the Keras SequentialAPI. For the binary classification of the 'infected' and 'notinfected', a dense layer with softmax activation was used and convolution layers with max-pooling were employed for feature extraction. The CNN model was compiled with Adam optimiser and binary cross-entropy loss function for monitoring accuracy during training. For the model training, 'early stopping' was implemented to prevent overfitting. The process had 20 epochs to strike a balance between the performance of the model and computational efficiency.

After the training, the model was tested on the validation dataset. Loss and accuracy of the model were evaluated which turned out to be [0.005764089524745941, 0.9973958134651184] respectively.

3.2.2 ResNet50 Model

eing renowned for proficiency in feature extraction from diverse datasets, pre-trained ResNet50 with pre-trained weights from the ImageNet dataset is chosen as the foundational architecture.

To control the features learned by the model, the layers of the base model are frozen to avoid alteration during the subsequent training. The pre-trained layers are frozen to prevent their weights from being updated during the subsequent training on our PCOS dataset. This ensures the retention of valuable features learned from ImageNet.

The custom model is built on top of ResNet50 utilising the global average pooling, dense layers, and dropout for effective extraction of features and classification. The final layer outputs probabilities for the two classes (infected, not infected) using softmax activation.

The model was constructed using the Adam optimizer with a learning rate of 0.001, categorical cross-entropy loss (suitable for the binary classification task), and accuracy as the evaluation metric. Data augmentation was done to increase the variability of the training dataset.

Generators are used to organise the training and validation datasets. The class mode is set to categorical for multi-class classification, and the target size ensures uniform dimensions for the input images.

The model was trained using the fit method, taking advantage of the generators for both training and validation. Early stopping was implemented to halt training if there was no improvement in the validation loss.

We evaluated the model on the validation set and visualized the training history through loss plots. Further, we conducted a detailed analysis using a confusion matrix, classification report, and ROC curve with AUC to comprehensively assess the model's performance. The loss and accuracy of the model were evaluated which turned out to be [0.12065348774194717, 0.9505208134651184] respectively.

3.3 Integration

n combining the results predicted from both the profiles, i.e. the blood profiles (taking the ExtraTrees Classifier model as the final model for taking the input) and the USG report (taking the CNN model as the final model for taking input), as discussed with the medical practitioners, if both the modules predict 'Yes' then there are high chances that the person has PCOS. If both the modules predict 'No' then chances are almost negligible that the female has PCOS and can continue with her current lifestyle. If either of the modules predicts 'Yes' but the other one predicts a 'No' then there are chances that she might have PCOS so for confirmation the woman shall go for further blood tests which include the tests like testosterone level, androgen level, LH-FSH, TSH, fasting glucose, HbA1c, and cortisol.

4 Performance Analysis

To evaluate the effectiveness and efficiency of the two modules, various metrics like accuracy, precision, recall, ROC curve, and precision-recall curve have been used for

the blood profile analysis. Whereas, for the ultrasound image analysis Loss vs epochs curve, accuracy vs epochs curve, and accuracy have been used as an integral part of the testing. The explanation of the parameters is:

4.1 Accuracy

It is a measure of how often a model correctly predicts the class of an instance.

$$Accuracy = \frac{TruePositives + TrueNegatives}{TotalInstances} \tag{1}$$

4.2 Precision

Precision measures the accuracy of positive predictions made by the model. It is the ratio of true positives to the sum of true positives and false positives.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2}$$

4.3 Recall (Sensitivity or True Positive Rate)

Recall measures the ability of the model to capture all the relevant instances of a class. It is the ratio of true positives to the sum of true positives and false negatives.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3}$$

Table 1 Values of the Parameters used in blood profile analysis

Algorithm	Accuracy	Precision	Recall
Logistic regression	0.853	0.786	0.688
Decision Tree	0.835	0.719	0.719
Voting Classifier	0.862	0.774	0.75
Bagging Classifier	0.872	0.8	0.75
Random Forest Classifier	0.872	0.8	0.75
AdaBoost Classifier	0.872	0.8	0.75
XGBoost Classifier	0.817	0.676	0.719
KNN	0.284	0.287	0.969
Gaussian Naive Bayes	0.798	0.614	0.844
Linear Discriminant Analysis	0.881	0.88	0.688
Quadratic Discriminant Analysis	0.761	0.56	0.875
MLP Classifier	0.872	0.846	0.688
ExtraTrees Classifier	0.998	1.000	0.994

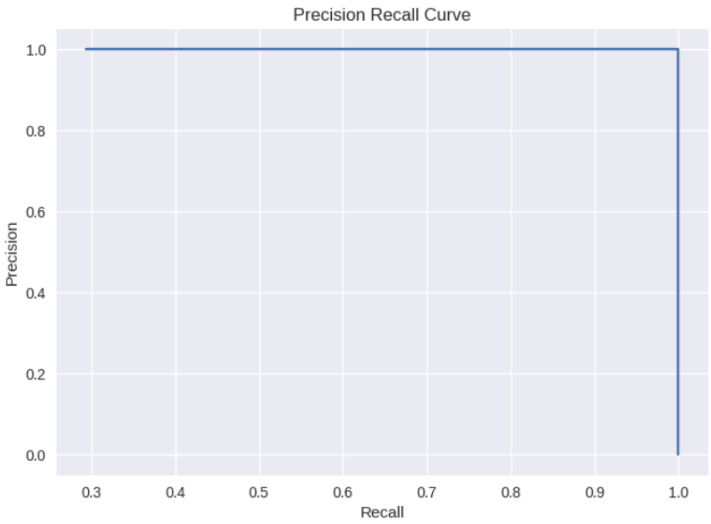

Fig. 4 Precision-Recall curve for the best model used in Blood Profile Analysis

4.4 Precision-Recall Curve

This type of curve is very useful when we have an imbalanced class distribution where one class is significantly more prevalent than the other one. The PR curve is created by plotting precision against recall for different threshold values.

4.5 ROC Curve

The ROC curve is a graphical representation illustrating the balance between the true positive rate aka sensitivity and the false positive rate that is 1-specificity for several different threshold values.

4.6 Loss vs Epochs Curve

?? This plot displays how the loss changes during the training of the model across multiple epochs. It helps in understanding how the model is learning from the training data and how its performance evolves with time. The downward going or descending curve indicates that the model is learning and adapting to the training data being fed to it. Overfitting can be determined easily if there decrease in the training data but a significant increase in the validation data. An increase in loss signifies that there may be convergence issues or insufficient complexity of the model.

4.7 Accuracy vs Epochs Curve

It determines how the accuracy of the model changes during the training of the model with respect to multiple epochs. An increasing curve indicates that the model is adapting and learning to correct its predictions. Overfitting can be checked if there is a significant increase in training accuracy but poor performance on the validation set. Fluctuations tell us that there are convergence issues.

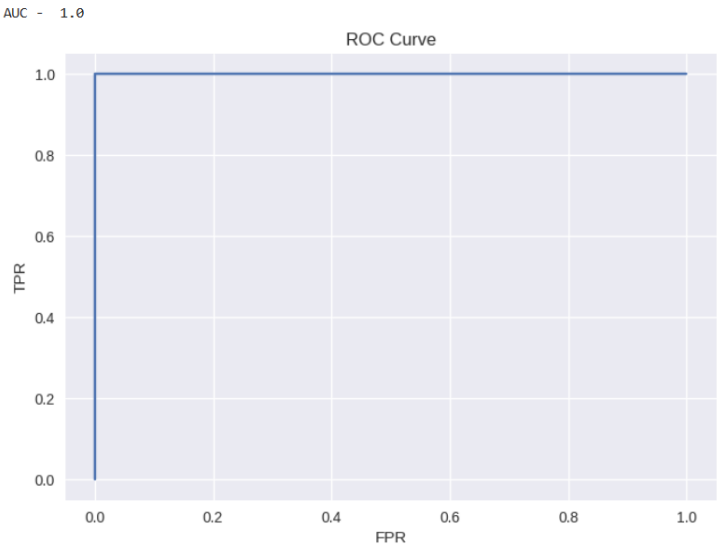

Fig. 5 ROC curve for the best model used in Blood Profile Analysis

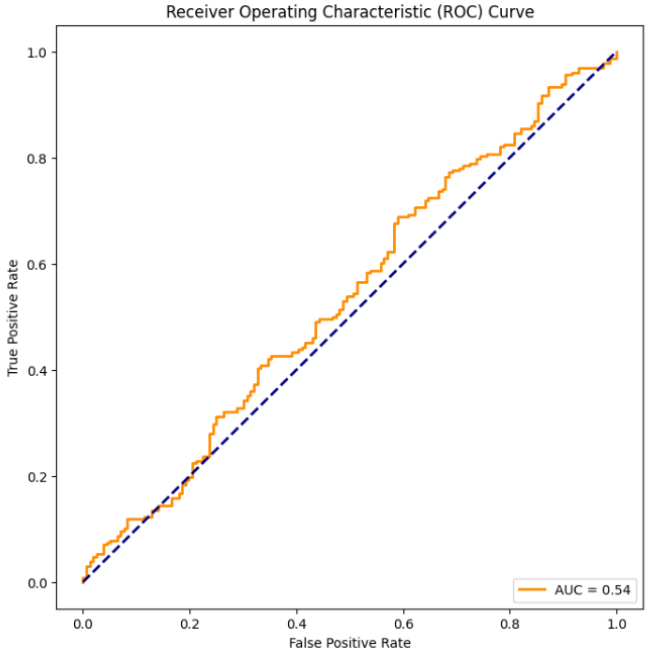

Fig. 6 ROC curve for the ResNet50 model used in Ultrasound Image Analysis

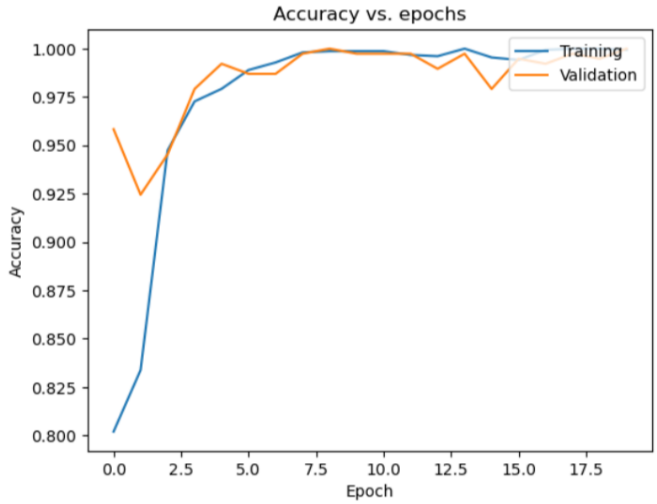

Fig. 7 Accuracy vs Epochs curve for the CNN model in ultrasound image analysis

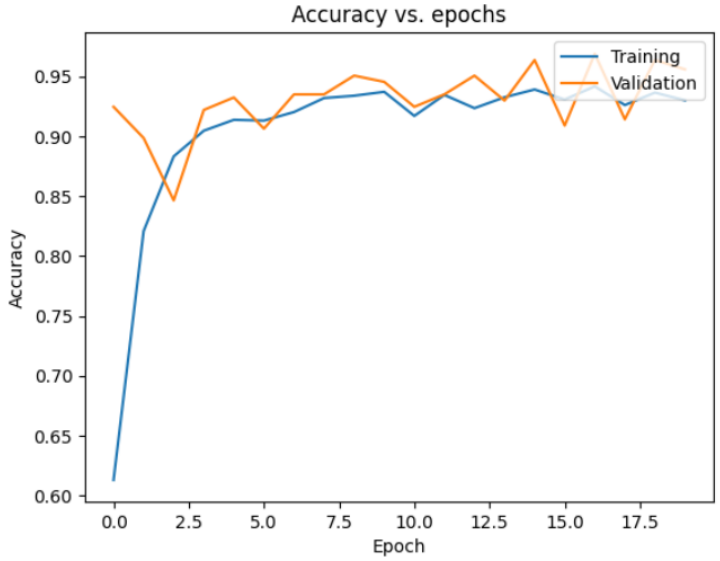

Fig. 8 Accuracy vs Epochs curve for the ResNet50 model in ultrasound image analysis

5 Conclusion

In this paper, we proposed an enhanced detection system for PCOS using the blood parameters as well as the ultrasound images of various patients. As it is a common condition faced by women in their reproductive age we manifest to help them as much as possible. Utilising various machine learning algorithms helped us gain insights

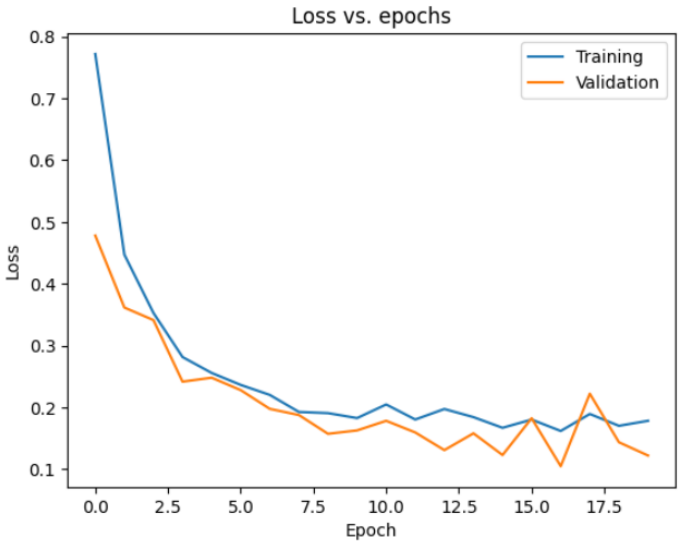

Fig. 9 Loss vs Epochs curve for the ResNet50 model in ultrasound image analysis

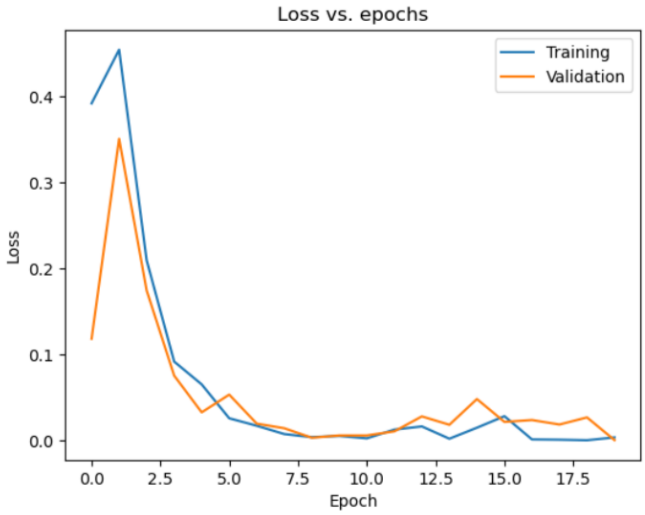

Fig. 10 Loss vs Epochs curve for the CNN model in ultrasound image analysis

into various patterns as how hormonal imbalances take place in reproductive age in a woman's body. The ultrasound image analysis gave us insights into the ovarian morphology telling more about how the structure and features of the ovary and follicles around it can cause this syndrome. By this dual-modality diagnostic approach, we tried to reinforce better accuracy and reliability of PCOS detection.

Table 2 Summary of the CNN model used in the USG image analysis

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, None, None, 12)	912
max_pooling2d (MaxPooling2D)	(None, None, None, 12)	0
conv2d_1 (Conv2D)	(None, None, None, 8)	2408
max_pooling2d_1 (MaxPooling2D)	(None, None, None, 8)	0
conv2d_2 (Conv2D)	(None, None, None, 4)	804
max_pooling2d_2 (MaxPooling2D)	(None, None, None, 4)	0
flatten (Flatten)	(None, None)	0
dense (Dense)	(None, 2)	6274

5.1 Future Work

In the future, this work can be potentially made better if there exist more blood tests in the blood reports (including the others already present) like 'HbA1c', 'cortisol', 'testosterone level, 'androgen level', 'LH', 'FSH' etc. The study can be made even more accurate if there is availability of combined datasets of patients which have both the blood tests including the tests mentioned above and their USG report to get better insights. This would thus escalate the rate of detection and also make it more sensitive and specific.

6 Acknowledgement and Declaration

6.1 Acknowledgement

The authors thank the School of Computer Science and Engineering, Vellore Institute of Technology, Chennai for spending their valuable time and efforts in sharing their knowledge and for helping us in every aspect.

6.2 Authors Contribution

All authors contributed to the study, conception and design. Material preparation and analysis were performed by Shatakshi Shree, Sandosh S. The first draft of the manuscript was written by Shatakshi Shree, and all the authors commented on previous versions of the manuscript. All the authors read and approved the final manuscript.

6.3 Funding

No funding was received in any form from any organisation for the submitted work.

6.4 Availability of data and materials

The datasets used and/or analyzed during the current study are available openly on Kaggle.

6.5 Conflict of Interest

The authors have no conflict of interests to declare that are relevant to the contents of this article.

References

[1] Denny, A., Raj, A., Ashok, A., Ram, C.M., George, R.: i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques. In: TENCON 2019-2019 IEEE Region 10 Conference (TENCON), pp. 673–678 (2019). IEEE

[2] Bharati, S., Podder, P., Mondal, M.R.H.: Diagnosis of polycystic ovary syndrome using machine learning algorithms. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1486–1489 (2020). IEEE

[3] Nandipati, S., Ying, C., Wah, K.K.: Polycystic ovarian syndrome (pcos) classification and feature selection by machine learning techniques. Appl. Math. Comput. Intell **9**, 65–74 (2020)

[4] Mehrotra, P., Chatterjee, J., Chakraborty, C., Ghoshdastidar, B., Ghoshdastidar, S.: Automated screening of polycystic ovary syndrome using machine learning techniques. In: 2011 Annual IEEE India Conference, pp. 1–5 (2011). IEEE

[5] Zad, Z., Jiang, V.S., Wolf, A.T., Wang, T., Cheng, J.J., Paschalidis, I.C., Mahalingaiah, S.: Predicting polycystic ovary syndrome (pcos) with machine learning algorithms from electronic health records. medRxiv, 2023–07 (2023)

[6] Thakre, V., Vedpathak, S., Thakre, K., Sonawani, S.: Pcocare: Pcos detection and prediction using machine learning algorithms. Biosci Biotechnol Res Commun **13**(14), 240–244 (2020)

[7] Danaei Mehr, H., Polat, H.: Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. Health and Technology **12**(1), 137–150 (2022)

[8] Silva, I., Ferreira, C., Costa, L., Sóter, M., Carvalho, L., C. Albuquerque, J., Sales, M., Candido, A., Reis, F., Veloso, A., et al.: Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models. Journal of Endocrinological Investigation, 1–9 (2022)

[9] 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), 32 (2023)

[10] Nasim, S., Almutairi, M.S., Munir, K., Raza, A., Younas, F.: A novel approach for polycystic ovary syndrome prediction using machine learning in bioinformatics.

IEEE Access **10**, 97610–97624 (2022)

[11] Aggarwal, S., Pandey, K.: Early identification of pcos with commonly known diseases: obesity, diabetes, high blood pressure and heart disease using machine learning techniques. Expert Systems with Applications **217**, 119532 (2023)

[12] Jaralba, J.R., Baldovino, R., Co, H.: A machine learning approach for initial screening of polycystic ovarian syndrome (pcos). In: Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications, pp. 517–529 (2021). Springer

[13] Prapty, A.S., Shitu, T.T.: An efficient decision tree establishment and performance analysis with different machine learning approaches on polycystic ovary syndrome. In: 2020 23rd International Conference on Computer and Information Technology (ICCIT), pp. 1–5 (2020). IEEE

[14] Lim, J., Li, J., Feng, X., Feng, L., Xia, Y., Xiao, X., Wang, Y., Xu, Z.: Machine learning classification of polycystic ovary syndrome based on radial pulse wave analysis. BMC Complementary Medicine and Therapies **23**(1), 409 (2023)

[15] Katarya, R., Jindal, A., Duggal, A., Shah, A.: A novel polycystic ovarian syndrome diagnostic system using machine learning. In: Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020, pp. 555–563 (2021). Springer