Science Eve Publisher

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

*Shatakshi Shree 1, Sandosh, S.1**

1 School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, Tamilnadu, India.

Corresponding author: sandosh.s@vit.ac.in*

ABSTRACT

Introduction Polycystic ovary syndrome (PCOS) is a common endocrine disorder in women of reproductive age associated with hormonal disturbance. The imbalance leads to many problems like irregular periods, pimples, hair loss, thinning of hair and in many cases cramps as well will also lead to problems ovaries where the person develop insulin resistance and metabolic syndrome which finally leads to infertility or unstable mood. Although PCOS is common, many cases go undiagnosed and PCOS has recently been considered as a psychosomatic disorder. This is a holistic study of PCOS diagnosis with a combination method by blood parameters and ultrasound images. For blood based diagnostics: Machine learning algorithms such as random forest with hyperparameter tuning through RandomizedSearchCV and GridSearchCV, Decision Tree of cost complexity pruning, Logistic regression, KNN, NaiveBayes, Neural network, LDA, QDA, Nearest centroid classifier, process classifier(Full grown tree),Voting classifier(lr + dt+ svm), Bagging classifier Classifier (etree), (Adaboost classifier), XGBoost. It uncovered patterns that highlighted the importance of hormonal imbalances as well as 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 multi-faced 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. Such approaches, although informative, may prove insufficiently specific and sensitive for accurate diagnosis in clinical settings and may consequently result in mis-diagnosis or delayed interventions. Furthermore, relying exclusively on blood profiles neglects those structural components of ovarian morphology that are essential for a precise diagnosis of PCOS. Meanwhile, ultrasound imaging is a useful method to achieve an overall view of ovarian morphology and cysts. Yet, while it alone is important for detecting ovarian endometriomas or cysts, ultrasound alone may miss subtle differences in other features of the ovaries and may give incomplete information. Thus PCOS is considered not only a disease rather it can also be claimed as a lifestyle disorder that can manage with diet and active life. The very standards of care demand a more subtle and comprehensive diagnostic process in this wedding. Thus, the prospect of hybrid models integrating blood profile analyses and ultrasound imaging has appeared.285 Advancements in machine learning284,286 and image processing287 —as well as artificial intelligence272—constitute a timely opportunity for making such models viable. Being incurable is an issue however so it is leading to life. Though it can not be cured but by taking proper medication and live a healthy life, one can manage their condition but there is an urgent need for improved complex models which augment the accuracy and effectiveness of the detection of PCOS with integrated health care efforts targeting precision medicine to secure early diagnosis with effective treatment. One goal of this research is to solve this problem by introducing a new hybrid method that tackles blood profiles and ultrasound image analysis for complete method of polystic ovarian syndrome (PCOS) diagnosis.

This study has twofold novel contributions. The research addresses the limitation of previous methods and uses data up to October 2023 for diverse and representative dataset using machine learning algorithms like(1) Random Forest, (2) Logistic regression,(3) Decision Trees, (4)Voting classifier, (5) Bagging classifier, (6) Extra trees (7) Adaboost classifier (8) KNN (9) MLP (10) Gaussian Na¨ıve Bayes model (11). By synergistically combining the strengths of multiple models, the method aims to reach enhanced precision in detecting PCOS using blood profiles.

.

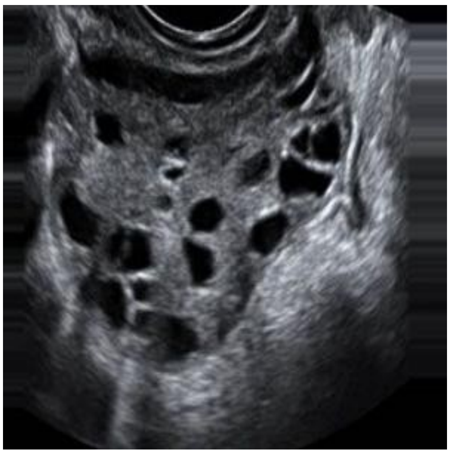

Fig. 1 USG of a patient with PCOS Fig. 2 USG of a healthy woman

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:

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 in-depth study improves the diagnostic ability of PCOS and also is important for precision medicine to be a mapping approach. The hybrid methodology aims to create comprehensive and refined indices of PCOS by aggregating information obtained from blood profiles with visual details observed on ultrasound imaging, potentially facilitating better clinical outcomes and enhancing personalization in the management of PCOS. This research attempts to identify the grey areas between conventional diagnostic and advanced technology, while seeking to marry existing constructs of PCOS detection mechanisms. The hybrid method is to combine the complementary information contained in the blood profiles with the ultrasound images, serving as a stronger and more comprehensive diagnostic framework for PES. Why is this study important: This study is important as it aims to provide another avenue (biomarkers) for the diagnosis of PCOS by bypassing some of the limitations posed by current diagnostic methodology.

Most of the rest of existing studies on detecting PCOS majorly rely on either biochemical markers or ultrasound image data separately. Although a number of models have shown impressive performance by utilizing either machine or deep learning algorithms, only few efforts have been made to introduce a comprehensive framework that combines the physiological and morphological evidence into diagnosis. In addition, many of the current approaches focus exclusively on predictive accuracy without accounting for scenarios in which one diagnostic modality returns an indeterminate result. This study addresses these gaps by proposing a hybrid framework for PCOS detection using complementary diagnostic evidence from blood-profile analysis and ultrasound imaging.

The main contributions of this research work are as follows:

- Created a predictive machine learning model with high accuracy and reliability for predicting PCOS using blood report data of metabolic and clinical attributes.
- Offered a holistic diagnosis model which integrates predictions made by blood-based and USG-based modules to provide an accurate and unbiased management solution.

The document is structured into three sections as follows: Section 2 offers a summary of the related studies. Section 3 gives detailed relative Background of the proposed framework and algorithms for Blood Profile analysis and USG analysis. Section 4 demonstrates analysis of the results of proposed work. Finally, Section 5 concludes with a summary of the results and an indication towards potential future work in this area.

2 Related Work

As for the study [1], which analyze 541 cases, there were 364 normal and non-PCOS {normal controls (NC) were composed of ultrasonically confirmed healthy subjects}and PCOS than 177. Eight potential features are selected after PCA that is discriminative between PCOS and non-PCOS. The Random Forest Classifier is our best model with an 89% accuracy on unseen data. These findings highlight the inadequacy of biochemical profiles or ultrasound results in isolation at diagnosing PCOS. AntiMu¨llerian Hormone (AMH) has emerged as a potential marker for PCOS detection. The accuracy of this study (89%) is a bit worse than other papers [as shown in previous research studies]; optimization by classifier weight parameters is being planned Clarification; Limitations Professional rotating pedometer and classification algorithms also need more expressive features.

In this research paper [2], the authors propose computer aided detection of women with PCOS/PCOD. The authors have fed data on 541 women and their 43 attributes from Kaggle data. It also implements a uni-variate feature selection method to find the features that predict PCOS the most for each class. They reached the conclusion that ratio of both hormones is most important and hence, they focused their study on these two factors combined. Various ML algorithms are tried and tested like gradient boosting, random forest, logistic regression and Random Forest Logistic Regression (RFLR). Among applied methods, RFLR proved best and demonstrated testing accuracy of 91.01% along with recall value of 90%, by harnessing a technique called as 40-fold cross-validation on top 10 features. It emphasizes the need for early screening and thus points out at better choice of features, makes it more (and less time taking) accurate. They validate their results against other studies and demonstrate that the RFLR method performs robustly when only 10 features are input, leading to faster processing and conservation of computational resources.

The research paper [3] addresses the prominent issue of Polycystic Ovarian Syndrome (PCOS), being an endocrine disorder in 5–10% of females in their adolescent age. Objective of this study is to enhance early diagnosis and treatment by using RapidMiner for model selection and feature encoding on Kaggle dataset. The Random Forest classifies the overall data (93.12%, RapidMiner) as high precision with Random Forest as its accuracy method. Some results are close to each other; in this case, KNN and SVM work comparative (90.83% at 10 selected features, RapidMiner). If we focus on feature selection analysis states that Competitive Results

(10-f) and Significance Feature (24) are two representations of competitive ground truth using; it can be found that 10 features yield quite respectable performance when the exact nature of gold standard is unknown while 24, with the most distinctly different variance (by at least an order of magnitude). The study then compares the accuracy, precision and recall of Python and RapidMiner and it shows that RapidMiner does a better job. Nonetheless, the paper recognizes this is dependent on the dataset type and methods implemented. In conclusion, this study demonstrates that machine learning is a potentially useful tool for the prediction of PCOS and demonstrates that careful selection of algorithms according to the characteristics of datasets will enable statical sound assessments in future studies.

This sliding window method is the main reason why this research paper [4] can show how do we developed an automatic PCOS detection method with clinical and metabolic parameters as early markers of the disease. Statistically significant features based on the parameters are used to construct a feature vector for an algorithm being proposed through a two-sample t-test. The classifiers use Bayesian and Logistic Regression for classification 93.93% accuracy is achieved with a Bayesian classifier while 91.04% accuracy with logistic regression. The automated system demonstrates its capability for use as an assistance to doctors, helping them save time on patient examinations and helping in early diagnosis of PCOS risk [7]. Furthermore, the study provides insights into clinical characteristics such as FSH, LH, BMI, and cycle length to differentiate normal from PCOS groups. In conclusion, this study proves that the Bayesian classifier is efficient for the early detection of PCOS and better than logistic regression, while promising higher accuracy through new classifiers.

The research highlights the importance of better predicting Polycystic Ovary Syndrome (PCOS) in an early stage, particularly within a larger group of outpatients with documented markers for this condition without their having been diagnosed [5]. The modelling study utilized data from EHR of 30,601 women treated at a SafetyNet hospital and created predictive models via logistic regression, support vector machine (SVM), gradient boosted trees and random forests. The models used hormone values and other clinical parameters to predict PCOS outcomes. The area under the curve (AUC) for out-of-sample test sets revealed promising sensitivity and specificity, ranging from 80% -85%. Hormones/adiposity phenotypes indicated higher likelihoods whilst pregnancy number and bHCG positivity indicated lower likelihood basing on ROC curves. We emphasize tools incorporated into EHR for early detection, counseling, and intervention to avoid long-term sequelae of excessive risk factors associated with PCOS. However, the identification of models requiring further external validation in different hospital-based populations is needed.

Research [6] Polycystic Ovary Syndrome (PCOS) is one of the most menacing diseases for women aged 18 to 44 years because of hormone imbalances which may lead to46 Irregular menstrual cycles,47 Excessive weight gain, Facial hair5088 growth4899 Acne and10 Irregular37 menstruation11 Severe cases as12 they cause13 infertility. However, the current diagnostic and therapeutic strategies are ineffectual in early diagnosis and prognosis prediction. To mitigate this, we proposed a system using five machine learning classifiers. They used the CHI SQUARE method to select 30 out of 41 features in the dataset which are basis of feature vector. When we compared the results of each classifier, Random Forest emerge as the most accurate and reliable. Using our system was learnt iron this dataset which is owned by Prasoon Kottarathil (link to dataset in kaggle) and has thus the best positive results with PCOS diagnosis and prediction.

The study [7] Polycystic ovary syndrome (PCOS) is a common metabolic and hormonal disorder in women of reproductive age that has compelling and renowned implications on their health, emphasizing the need for earliest diagnosis for optimal management. Machine learning methods for medical diagnosis have been achieved with encouraging performance in recent years. Thirdly, feature selection method focuses on selecting the most relevant set of features which improve the computational efficiency and classifier performance. This explores the application of both classic and ensemble machine learning models on kaggle PCOS dataset; They use the dataset—that global retail dataset that has been used in the previous two articles (though, not all of its features)—to test some combined models through some selection methods for smaller sets of features. These results demonstrate that the right feature selection can have a huge impact on improving models. To our astonishment, however the Ensemble Random Forest model which Pipelines set with a limited characteristic via one of the methods outperformed the rest by far with an Accuracy of 98.89% and Sensitivity of 100%.

In their [8] work, they applied different machine learning (ML) algorithms to recognize important clinical and laboratory predictors related to diagnosis of Polycystic Ovary Syndrome (PCOS), as well as to cluster individuals into distinct phenotypic groups. It contains data from 72 patients with PCOS and 73 healthy women. We are using BorutaShap for prediction and Random Forest algorithm to segregate them. Lipid accumulation product (LAP), abdominal circumference, TAFI levels, BMI, CRP, HDLc, FSH, insulin levels and HOMA-IR value were singled out by the algorithm as pertinent for PCOS diagnosis among 58 factors tested; age also appeared as risk factor (overweight), along with prolactin and 17-OH progesterone followed by triglycerides levels and a first-degree relative history of diabetes mellitus. Using these features together, they obtain an accuracy of 86% and a ROC-AUC of 97%. The algorithmic process not only ranks important variables for PCOS but also clusters patients into statistically distinct phenotypes that may assist in the development of targeted and possibly more efficacious treatment protocols for PCOS.

This study [9] approaches the challenging disorder using AI by introducing heterogeneous ML and DL classifiers in predicting PCOS among fertile patients. The study applies a multi-stack of ML models using an open-source dataset from Kerala, India which contains a total of 541 patients and reports excellent performance with 98% accuracy, precision, recall F1-score: 97%, 98% respectively. Incorporating Explainable AI (XAI) techniques: one of the classical SHAP (SHapley Additive Values), and LIME (Local Interpretable Model Explainer), ELI5-Step1, Qlattice and with Random Forest step 2 feature importance(which is built-in step in many ML models) integrates transparency and improve interpretability of the model.

In this [10] study, considerable attention is given to the issue of Polycystic Ovary Syndrome (PCOS/PCOD) in reproductive aged women (10-12 years) and seeks to develop an ML-based predictive model for PCOS. PCOS, or Polycystic Ovary Syndrome is a hormonal disorder that causes problems related to functioning of ovary and leads to various issues in the later years of life. CSPCOS mechanism and relative approach existing research that draws upon clinical and physical parameters of women (data set until October 2023) A total of 10 hyper-parameterized machine learning models are utilized to compare all the others, and results determined GNB with 100% accuracy, precision, recall, and F1-scores alongside a very low time computation of 0.002 seconds. According to the k-fold cross-validation of GNB also provides a high accuracy score, which was up to 100%. This model is constructed based on elements like prolactin, blood pressure, thyroid stimulating hormone, relative risk and pregnancy that are significant in predicting PCOS at an early stage. This study is firstly in response to the needs of the medical community who are working proactively to decrease miscarriages and direct timely interventions for women by detecting PCOS early.

This [11] is a major problem and this research deals with it as a complex hormonal and reproductive disorder that afflicts around 10 million women worldwide. It is thought to affect one in 5 women in India. PCOS is one such metabolic disorder which can lead to various complications and hence early diagnosis is warranted. It mainly identifies the diseases which will act as a predictor of PCOS and utilize these data in the early identification of PCOS. This research combines known disease datasets, conducts feature selection and supervised/unsupervised learning algorithms to find relevant attributes for predicting PCOS. Key messages * The results highlight significance of obesity, heart disease, hypertension and diabetes in the early ascertainment of PCOS. This study has contributed in (i) formulation of a new dataset, (ii) application of various learning algorithms and (iii) providing insights for medical practitioners to aid early diagnosis of PCOS by using minimum number of features. The article ends with item 6 on results, future work and references. The problem of early and accurate diagnosis of Polycystic Ovarian Syndrome (PCOS) which is a common gynecological disorder among women worldwide is helped by this research [12]. There is an urgent need to develop accurate diagnostic tools as PCOS and its comorbidities have a significant economic burden on society, particularly among adolescent women. The study uses data analysis of 23 non-invasive parameters to evaluate a total of 7 ML algorithms for their ability to screen PCOS. The dataset consists of clinical data from 540 patients in Kerala, India, of which 378 are used for training and 162 are set aside to test the classifier. The study proposes ML algorithms capable of screening PCOS without the need for an invasive test. Our study provides evidence for the potential of ML models to be used as appreciable screening tools for PCOS, achieving high sensitivity and good ROC-AUC performance.

Thus, the study [13] concludes that Polycystic Ovary Syndrome (PCOS) is a great threat to women's health and leaves behind lifelong damage. A lot of women suffers unknowingly or the condition does not realize in its early stage. A long-term management plan for PSOC is important as it can have serious effects on a woman's health, contributing to the risk of conditions such as metabolic syndrome, sleep apnea, depression and endometrial cancer besides it being a treatable cause of infertility. These consequences can be averted, however, with careful monitoring and detection as early as possible. This study uses different machine learning based techniques to develop a decisiontree model in an optimized manner. After a woman gets to know from doctor that she possesses PCOS, then through decision tree, she would be able to monitor the major observed changes in her body or hormone level. This equips her to decide when to seek medical help, which allows for 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 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.

3.1.2.1 Feature Engineering and Scaling

After data cleaning, numerical attributes were inspected for consistency and outliers. Since the selected ExtraTreesClassifier is tree-based and inherently insensitive to feature scaling, normalization was not mandatory for the final model. However, missing values were imputed using median substitution to preserve data distribution. Correlation analysis and exploratory feature assessment were used to identify clinically relevant predictors such as follicle count, menstrual irregularity, BMI, LH-FSH ratio, and hormone-related parameters. No synthetic feature generation was performed.

3.2 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.

3.2.1 Model Evaluation

The 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. The splitting is performed in a way that does not search for the best split which entails far less computer power than decision tree and random forest. Best Model with Hyperparameters: The best model was obtained with hyper parameters n estimators=600, max depth=20, n jobs=-1 and random state=0. The model was trained on the dataset provided and its accuracy recorded on the same as 99.81%. While this resulted in a model with impressive predictive power, one should be careful to not overinterpret such high performance. The dataset is based on one public data source and may have strong feature correlations that make classification easier. Overfitting was reduced by applying ensemble learning and hyperparameter optimizations, but independent external validation on clinical datasets is still required before being applied to real-world healthcare settings. Future research will explore stratified k-fold cross-validation and multi-centre validation in order to examine generalizability.

Furthermore, it is the ExtraTreesClassifier which provides optimal performance with an appealing, balanced trade-off between prediction accuracy, computation speed and robustness. ExtraTrees differs from Random Forest in that you select split thresholds randomly, another level of randomness means less variance which leads to better generalisation. Additionally, the tree-based ensemble models have feature importance measures which presents a foundation for clinical interpretation of the most impactful biochemical parameters. ExtraTrees outperformed boosting-based approaches like XgBoost while requiring less hyperparameter search.

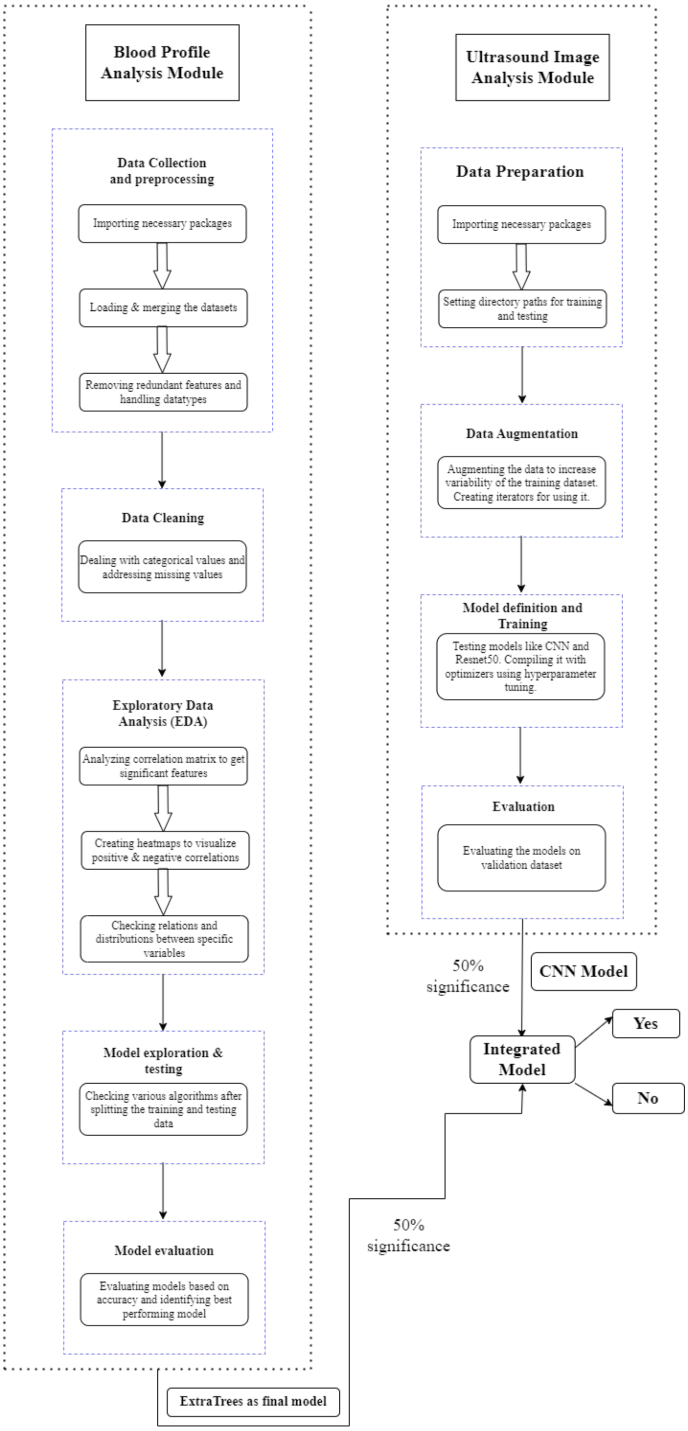

Fig. 3 Model Architecture of the Proposed Work

3.3. Ultrasound Imaging Analysis

In 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.

3.3.1 CNN Model

It 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 [0.005, 0.99] respectively.

The dataset was divided into training, validation, and testing subsets according to the folder structure provided in the source repository. Approximately 80% of images were used for training and 20% for testing, with a validation subset extracted from the training data during model development. Since the dataset was obtained from a public repository, explicit patient identifiers were unavailable. Therefore, patient-level separation could not be independently verified, and this is acknowledged as a limitation of the study.

Although transfer learning using ResNet50 demonstrated strong classification performance, the custom CNN model achieved higher validation accuracy and lower validation loss on the current dataset. The CNN architecture also contains fewer trainable parameters, resulting in reduced computational requirements and faster inference, making it more suitable for deployment in resource-constrained clinical environments.

3.3.2 ResNet50 Model

Ueing 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.

Training was performed using a GPU-enabled TensorFlow environment. The CNN model required fewer computational resources than ResNet50 because of its reduced architectural complexity. ResNet50 required longer training times owing to its deeper architecture and larger parameter space. However, both models remained feasible for offline clinical decision-support applications.

3.4 Integration

In 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.

The present study adopts a clinically interpretable rule-based fusion strategy derived from consultations with medical practitioners. While equal weighting was assigned to biochemical and ultrasound predictions for simplicity and interpretability, future work may investigate weighted fusion, Bayesian fusion, confidence-based decision aggregation, or meta-learning frameworks capable of dynamically assigning modality importance based on prediction certainty.

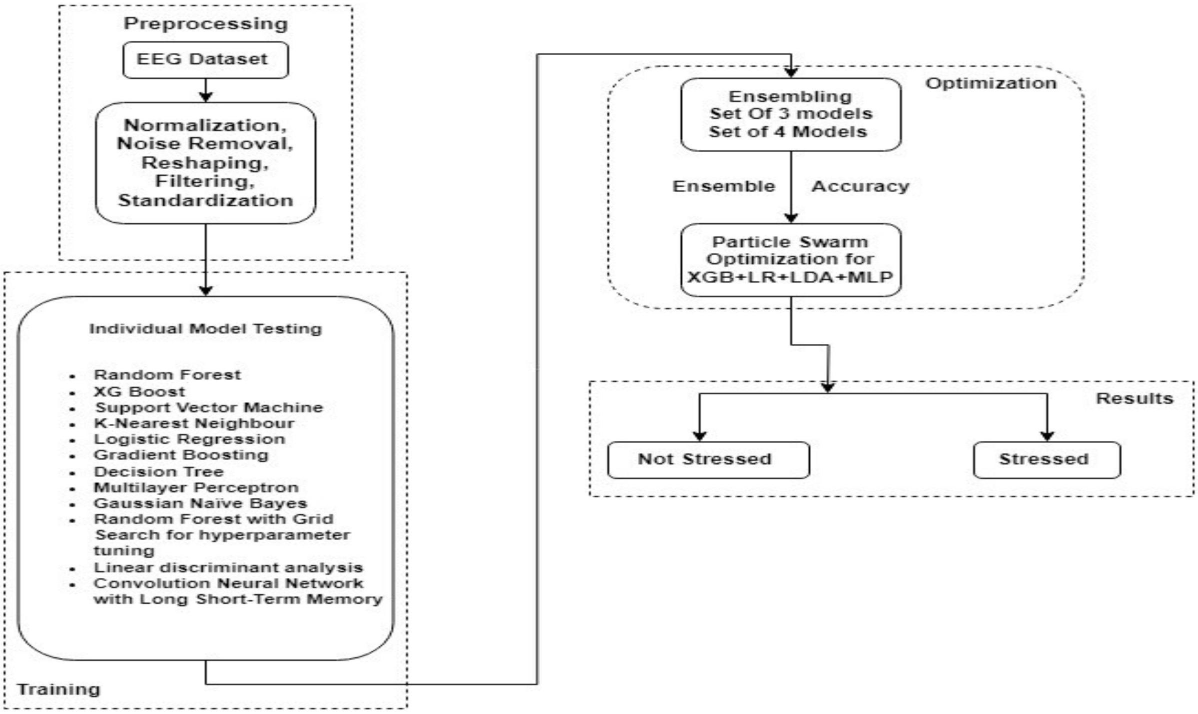

Fig 1. Overall Flow of the proposed technique

4 Experimental 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 forthe 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.

Although the reported performance metrics are encouraging, confidence intervals and statistical significance testing were not performed in the current study. Future work will incorporate bootstrapping and repeated cross-validation procedures to estimate confidence intervals and assess statistical significance between competing models.

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

Precision-Recall Curve

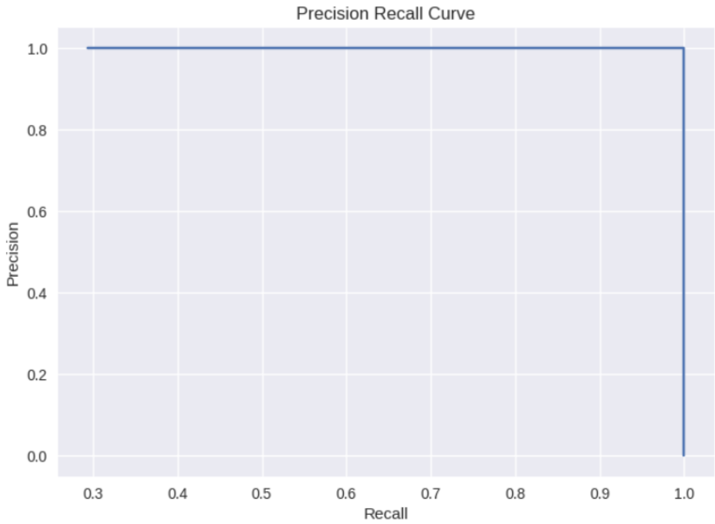

Fig. 4 Precision-Recall curve for the best model used in Blood Profile Analysis

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.

ROC Curve

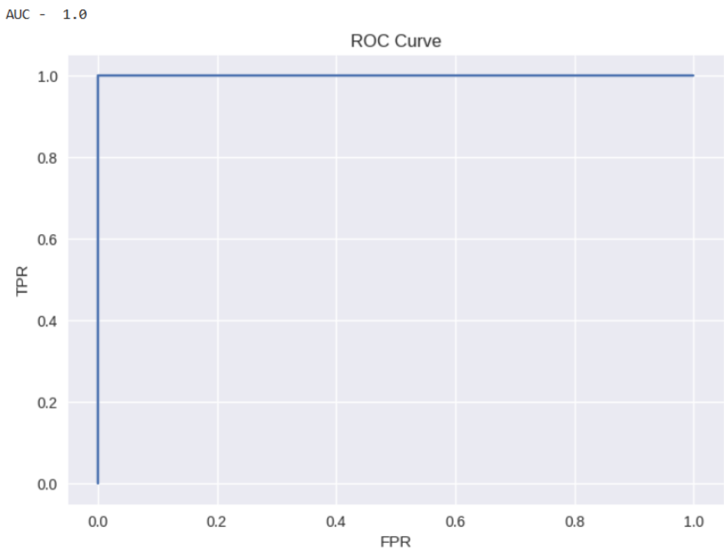

Fig. 5 ROC curve for the best model used in Blood Profile Analysis

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.

Loss vs Epochs Curve

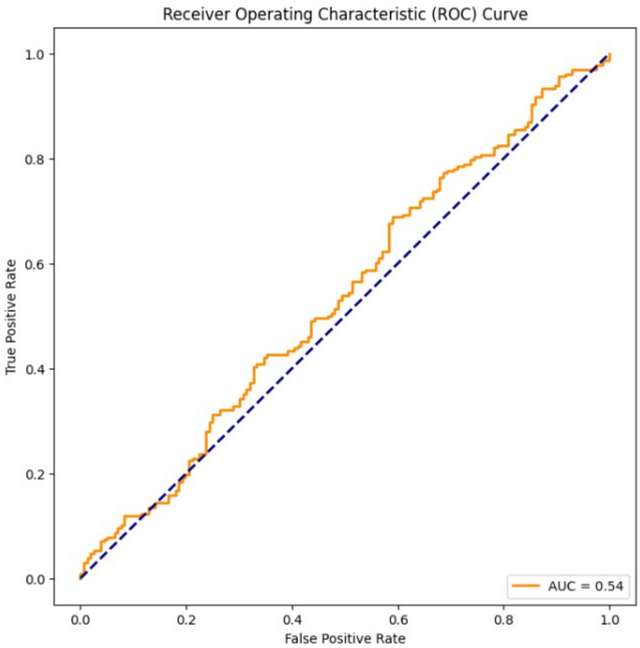

Fig. 6 ROC curve for the ResNet50 model used in Ultrasound Image Analysis

The 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.

Accuracy vs Epochs Curve

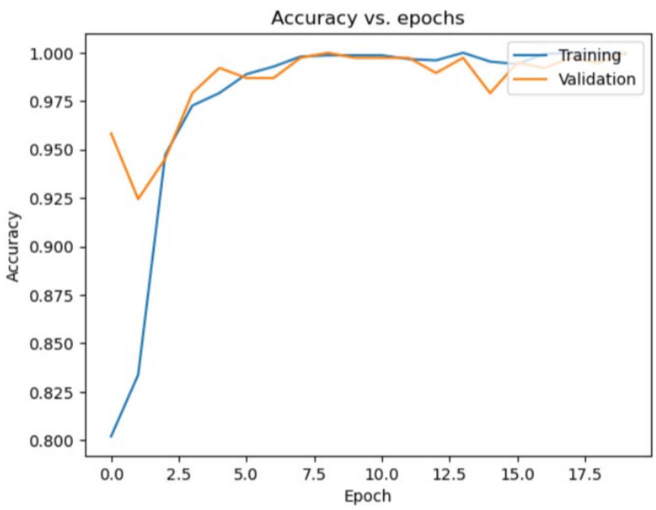

Fig. 7 Accuracy vs Epochs curve for the CNN model in ultrasound image analysis

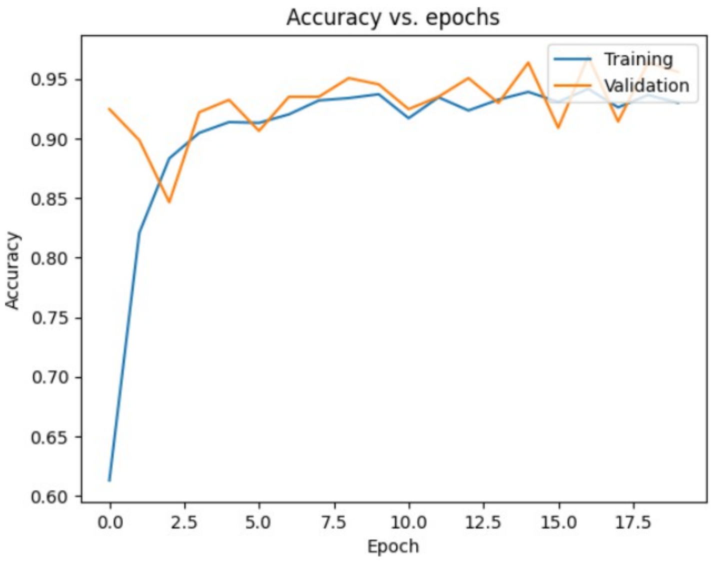

Fig. 8 Accuracy vs Epochs curve for the ResNet50 model in ultrasound image analysis

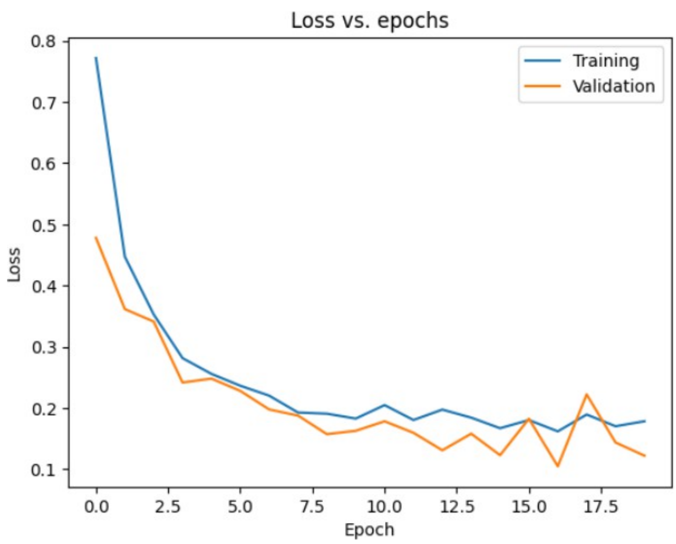

Fig. 9 Loss vs Epochs curve for the ResNet50 model in ultrasound image analysis

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.

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. Utilizing various machine learning 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.

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.

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., S´oter, 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