Expert Systems With Applications 203 (2022) 117592

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning

Shamik Tiwari [a,1], Lalit Kane [a,2], Deepika Koundal [a,3], Anurag Jain [a,*,4], Adi Alhudhaif [b,5], Kemal Polat [c,6], Atef Zaguia [d,7], Fayadh Alenezi [e], Sara A. Althubiti [f]

[a] School of Computer Sciences, University of Petroleum and Energy Studies, Dehradun 248007, Uttarakhand, India
[b] Department of Computer Science, College of Computer Engineering and Sciences in Al-kharj, Prince Sattam bin Abdulaziz University, P.O. Box 151, Al-Kharj 11942, Saudi Arabia
[c] Department of Electrical and Electronics Engineering, Faculty of Engineering, Bolu Abant Izzet Baysal University, 14280 Bolu, Turkey
[d] Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. BOX 11099, Taif 21944, Saudi Arabia
[e] Department of Electrical Engineering, College of Engineering, Jouf University, 72238, Saudi Arabia
[f] Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al-Majmaah 11952, Saudi Arabia

ARTICLE INFO

Keywords:
Polycystic Ovary Syndrome
Smart diagnosis
Machine learning
Random Forest
Out of Bag error

ABSTRACT

Polycystic Ovary Syndrome (*PCOS*) is a hormonal disorder that affects a large percentage of women of reproductive age. *PCOS* causes imbalanced or delayed menstrual cycles and produces high levels of the male hormone. The ovaries may create a significant number of little fluid-filled sacs (follicles) yet fail to discharge eggs regularly. The actual cause of *PCOS* is uncertain. However, early exposure and curing, as well as weight loss, may lower the threat of long-term complications. This study focuses on *PCOS* diagnosis based on a clinical dataset supplied by Kottarathil, accessible via its Kaggle repository. Non -invasive screening parameters are used to evaluate a range of machine learning approaches for screening *PCOS* patients without the use of invasive diagnostics. According to the findings of the experiments, the Random Forest (RF) method outperforms the other prominent machine learning algorithms with an accuracy of 93.25%. Further, the out-of-bag (OOB) error is utilized for assessing the prediction performance of RF.

1. Introduction

Polycystic Ovary Syndrome (*PCOS/PCOD*) is a serious disease that affects the ovaries of females in their reproductive years (15–45). 5 to 10% of females in the reproductive age range suffer from this disease (Escobar-Morreale, 2018)). *PCOS* is caused by a sex hormone imbalance. An increase in the level of androgens (male hormone) in females causes cysts in the ovary. Gradually, these lumps expand, then obstruct the process of ovulation. This disruption of ovulation in women with PCOS lowers their chance of pregnancy (Meier, 2018). Women suffering from *PCOS* are also more likely to suffer from type 2 diabetes. Other common symptoms of *PCOS* are irregular menstruation, excessive bleeding during menstruation, acne, oily skin, weight gain, headaches, irritable and angry behavior, sleep disorder, and male-like hair growth on the chest, stomach, back, and face (Teede et al., 2018). *PCOS* is generally considered to be a lifestyle disease, but the exact reasons behind its occurrence are still not known. Although it is difficult to completely fix this problem, with proper exercise, a nutritious diet, and

* Corresponding author.
E-mail addresses: shamik.tiwari@ddn.upes.ac.in (S. Tiwari), lalit.kane@ddn.upes.ac.in (L. Kane), dkoundal@ddn.upes.ac.in (D. Koundal), anurag.jain@ddn.upes.ac.in (A. Jain), a.alhudhaif@psau.edu.sa (A. Alhudhaif), kpolat@ibu.edu.tr (K. Polat), zaguia.atef@tu.edu.sa (A. Zaguia), fshenezi@ju.edu.sa (F. Alenezi), s.althubiti@mu.edu.sa (S.A. Althubiti).
[1] 0000-0002-5987-7101
[2] 0000-0002-7305-1189
[3] 0000-0003-1688-8772
[4] 0000-0001-5155-022X
[5] 0000-0002-7201-6963
[6] 0000-0003-1840-9958
[7] 0000-0001-9519-3391

https://doi.org/10.1016/j.eswa.2022.117592
Received 3 January 2022; Received in revised form 11 May 2022; Accepted 12 May 2022
Available online 14 May 2022

maintaining a proper body mass index, the effects of *PCOS* can be mitigated. Most women are not aware of the disease until they undergo a pregnancy test (Louwers & Laven, 2020). This delayed diagnosis often exacerbates the disease's severity. There is no exact medical examination to diagnose *PCOS* y, though cysts can be discovered through sonography, androgen levels can be checked through blood tests, and body hair can be evaluated from physical examination (Azziz et al., 2016; Barber & Franks, 2021).

Artificial intelligence (AI) and its subfields are progressively gaining attraction in everyday life, and are expected to play a significant role in disease detection and treatment in upcoming years (Maadi et al., 2021). This study applies AI in the ill-defined space of PCOS detection. The high dimensionality of all the features in a woman's life that can cause PCOS, plus difficulty in analyzing sonography cyst imaging, makes traditional diagnosis complex and time-consuming. Thus this computational approach may very well revolutionize PCOS diagnosis and improve the lives of millions of women.

1.1. Motivation

Lives can be drastically improved through early-stage detection of such diseases as PCOS. Models based on machine learning or deep learning are well-suited for the analysis of sonography images (Kiruthika et al., 2020). These models can suggest diagnoses on *PCOS* based on clinical patient history (Hassan & Mirza, 2020). Some researchers have also suggest that *PCOS* may be a genetic disease, so this paper's modeling research can also contribute to RNA-seq and microarray data obtained from genes (Xie et al., 2020).

1.2. Contribution

This work tackles the prediction of *PCOS*. The key contributions of the proposed work are as follows:

- Defined a correlation-based feature selection methodology for the Polycystic Ovary Syndrome patient dataset.
- Developed optimal classifier and correlation thresholds by evaluating many different machine learning algorithms on the PCOS dataset.
- Discovered optimal parameter tuning for the finally-selected Random Forest classifier through considering out-of-bag error.

1.3. Outline

First, Section 2 comprehensively walks through the pertinent literature. Subsequently, Section 3 discusses the methodology followed and the machine learning schemes implemented for *PCOS* diagnosis using non-invasive parameters. The results obtained are presented and analyzed in Section 4 with the concluding remarks drawn from the empirical analysis in Section 5.

2. Literature review

Past research reinforces the necessity for effective *PCOS* diagnosis. Hart and Doherty, 2015, established that a distinctly higher number of women with *PCOS* were admitted to hospitals as opposed to those for menorrhagia, infertility, and miscarriage, and could require in vitro fertilization. Early diagnosis and treatment of *PCOS* were also consolidated by the study conducted by Shan et al. in (2015). A multivariate analysis of data collected via questionnaire revealed that risk factors for *PCOS* were bad mood, menstrual cycle disorder, diabetes, family medical history, menstrual irregularity, history of infertility, and lack of exercise. Several machine learning (ML) algorithms, their customizations, and deployments have been reported in the literature to target early *PCOS* detection. The following section navigates the most important works.

Driven by the desire to establish a cost-effective preliminary diagnosis of *PCOS* patients, Jaralba et al. in (2020) evaluated various ML algorithms for efficacy. This evaluation was carried out over 23 non-invasive parameters, which lowered test costs and ameliorated patient experience. The 540-sample dataset (collected at Kerala state hospitals in India) is divided into 70:30 training and testing portions. Models were implemented and assessed for precision, accuracy, specificity, sensitivity, ROC, and AUC scores. In terms of sensitivity and ROC, analyzed models demonstrated considerable promise as a screening tool for *PCOS*, with some models outperforming other proposed techniques using invasive procedures.

Bhardwaj and Tiwari in (2022) highlighted the importance of artificial intelligence and its subfields in the domain of healthcare, particularly in the identification of PCOS. Authors used Pearson correlation to select the most suitable features and employed different algorithms like support vector machine, multi-layer perceptron, XG boost, and random forest classifier to optimize of the model. They utilized a Kaggle dataset containing data from 541 female patients, among which 32% were suffering from PCOS without fertility. They claimed to achieve an accuracy level of 93% with a support vector machine model. Zigarelli et al. in (2022) also utilized the same dataset but with a different algorithm. They utilized the CatBoost algorithm for classification and were able to achieve an accuracy level of 90.1%.

Working towards a rigorous evaluation for *PCOS* diagnosis, Bharati et al. in (2020) employed an open dataset from Kaggle comprising the samples from 541 women, of which around 32% were diagnosed with *OS*. These samples were first probed for the best features using a univariate feature-selection algorithm, followed by a subsequent ranking of the shortlisted features. Further, the classifiers, namely, random forest, hybrid random forest and logistic regression (RFLR), logistic regression, and gradient boosting were run on the selected features. The RFLR classifier outperformed the rest by achieving an accuracy and recall of 91 and 90%, respectively, on rigorous 40-fold cross-validation on the ten highest-ranked features.

Danaei Mehr and Polat in (2022) discussed the significance of feature selection methods and ensemble classifier techniques while detecting PCOS at an early stage. With the dataset available on Kaggle, authors analyzed the performance of ensemble random forest, multi-layer perceptron, AdaBoost, and extra tree classifiers. Through a simulation study, authors showed the significance of the feature selection technique and also that the ensemble random forest algorithm gave the best performance with the reduced set of features.

Zhou et al. in (2018) observed that known *PCOS* genes were potentially orthogonal features due to certain distinguishing characteristics. The characteristic difference was driven by three observations of *PCOS* genes: proximity to the network center, mutual interactions, and enhanced discrimination capabilities under certain biological processes. The authors devised an ML procedure to forecast fresh *PCOS* genes that classified 233 *PCOS*-labelled candidates with a 90% posterior probability. The ML predictions were consolidated by 70% of the evidence. To discriminate between *PCOS* and *Non−PCOS* genes, the Kolmogorov–Smirnov (KS) test was applied. The Kolmogorov–Smirnov (KS) test differentiated the features between *PCOS* and *Non−PCOS* genes. Subsequently, a Support Vector Machine (SVM) classifier with a liner kernel was applied to generate the claimed results. To accomplish this, direct interaction neighbors were counted for each gene, with the observation that *PCOS* genes were higher in degree than their *Non−PCOS* counterparts.

Guleken et al. in (2022) used blood serum to identify Polycystic Ovary Syndrome through hormone levels. Authors have used Raman spectroscopy to monitor biomolecules in the serum. The authors collected samples of blood serum from a group of healthy women and another group of women suffering from PCOS. Authors combined chemometric measurement, spectroscopic measurement, and correlation analysis to differentiate the blood serum of healthy women and women suffering from *PCOS* with accuracy ranging from 80.92% to 96.04%. The

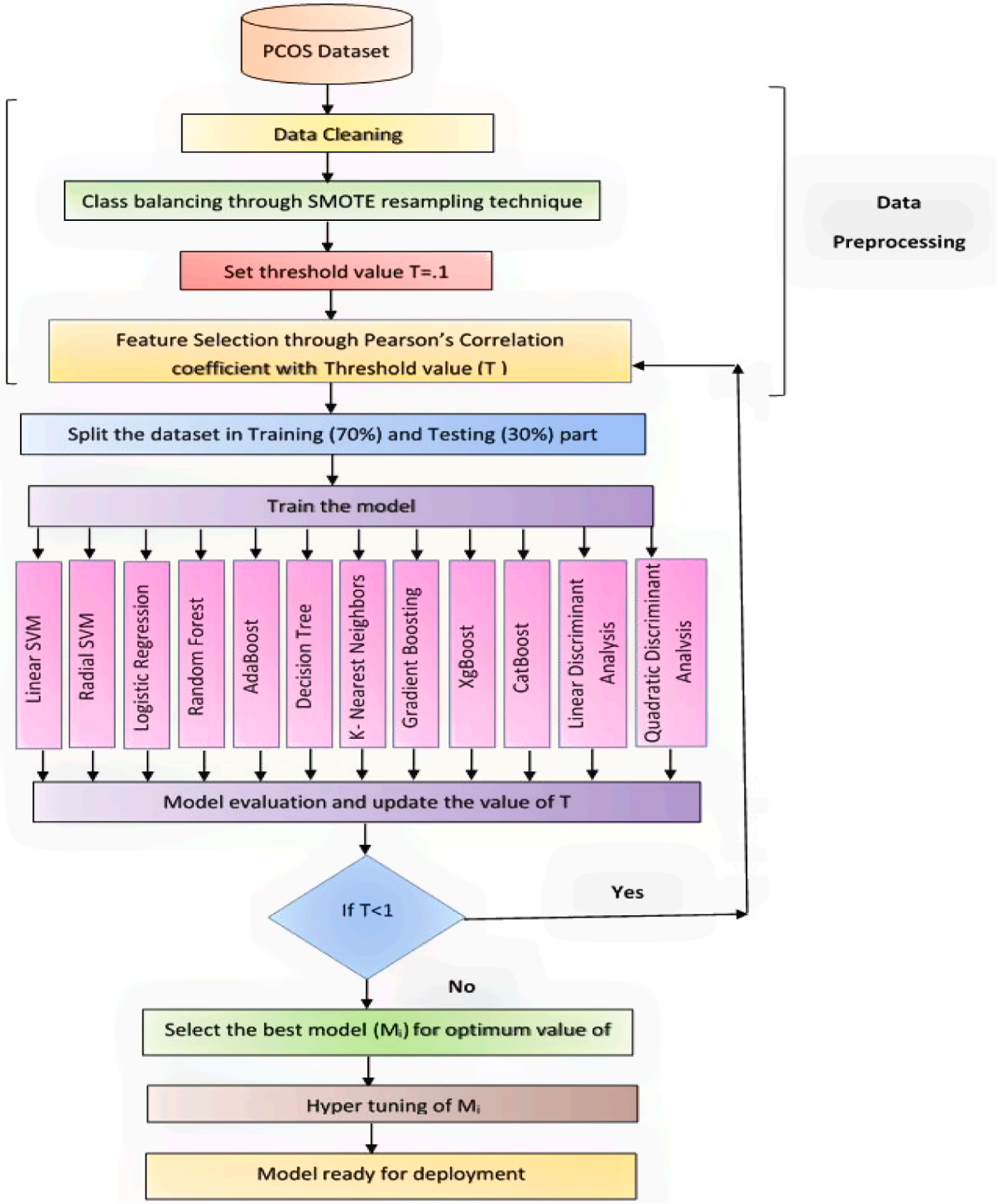

Fig. 1. Proposed methodology for PCOS diagnosis using machine-learning models.

variance in accuracy was due to the varying machine learning models. The maximum accuracy was achieved through the kNN model with the value of k as 3.

Mehrotra et al. in (2011) asserts that clinical and metabolic parameters act as preliminary diagnostic clues for the disease. The clue-based detection algorithm proposed by the authors selects features from these clues by ascertaining their statistical significance through t-tests. Bayesian and Logistic Regression (LR) classifiers were applied, and it was deduced that the Bayesian classification achieved superior results (94% as against 91% in terms of accuracy).

Bharati et al. in (2022) utilized the data-driven approach to diagnose *PCOS*. They used the Kaggle dataset to train and test the model. The authors also applied feature selection and elimination with cross-validation. Catboost, voting soft, and voting hard are the classifiers used by the model development. Their proposed model was able to achieve the highest accuracy of 91.12% with the soft voting classification model.

A questionnaire-based dataset collected by Vikas et al. in (2018) was evaluated through a set of established classification algorithms to assess its predictive information for *PCOS* detection. Sensitivity, specificity,

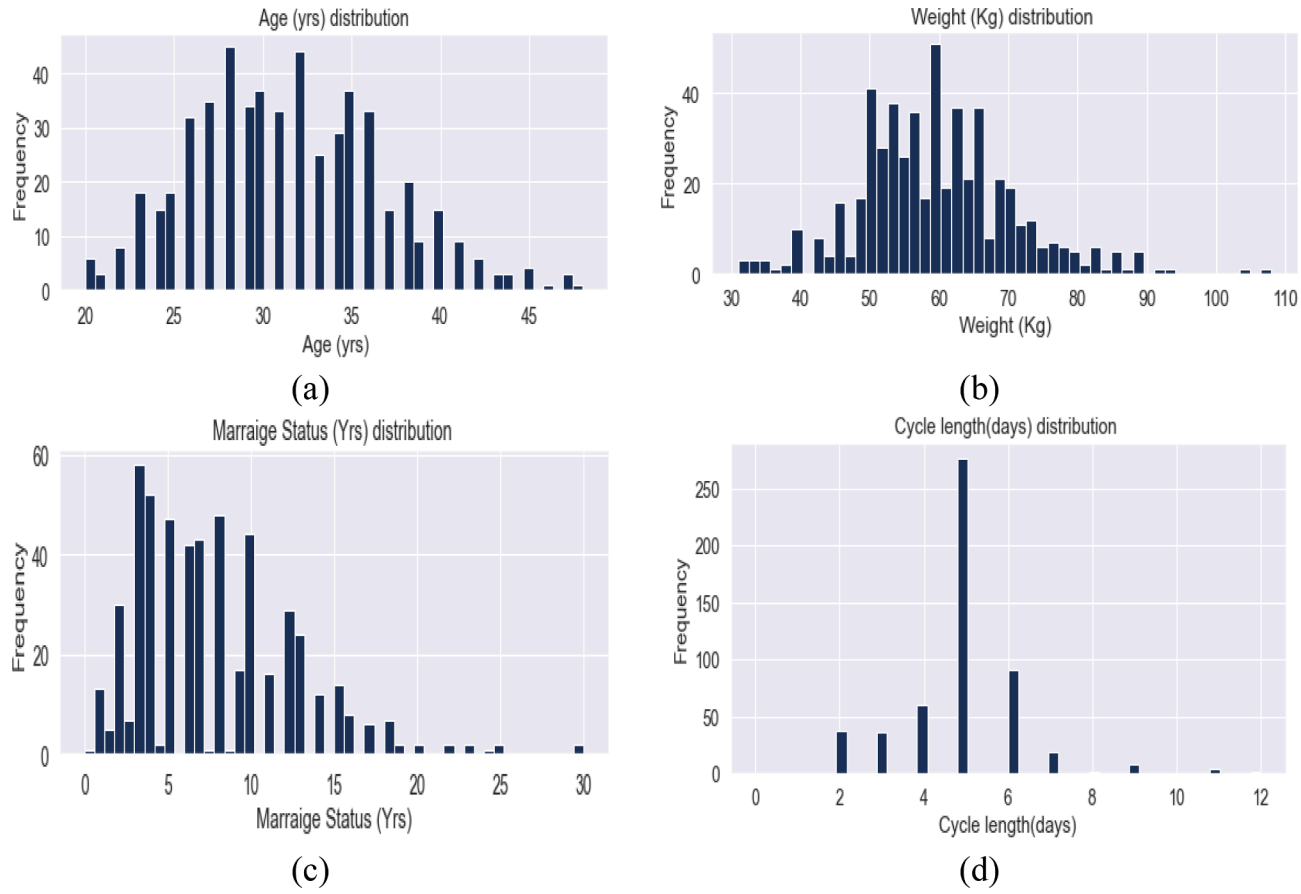

Fig. 2. Frequency distribution of age, weight, marriage status, and cycle length features from PCOS dataset.

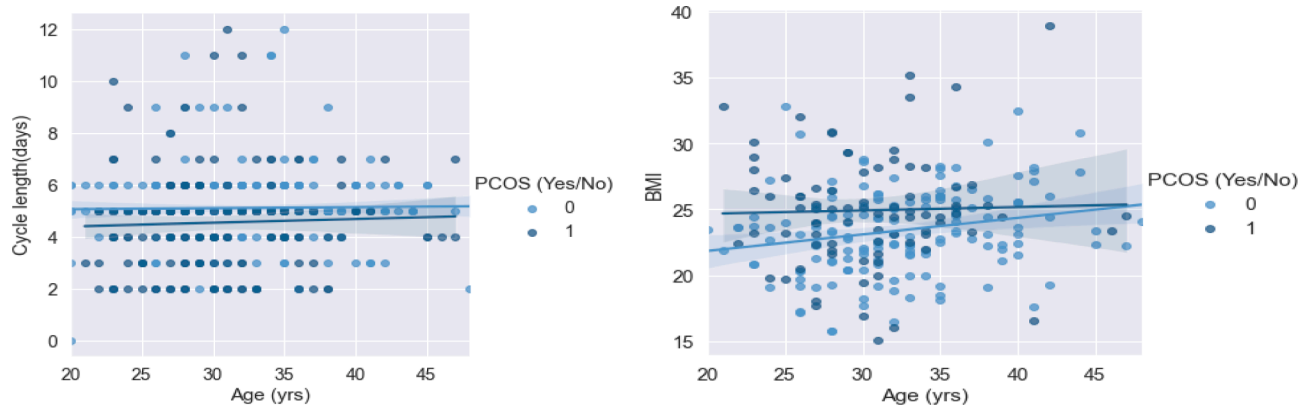

Fig. 3. (a) Visualizes the linear relationship as determined through regression for the length of the menstrual phase in PCOS vs normal, (b) Visualizes the linear relationship as determined through regression for a pattern of weight gain (BMI) over years in PCOS and Normal.

and precision were selected as the performance metrics. The Bayesian decision tree C 5.0, the backpropagation neural network, and naïve Bayesian were used to confirm the dataset's efficacy, in which the Bayesian one was marginally ahead of the others (97% accurate).

The goal of both Mehrotra et al. in (2011) and Vikas et al. in (2018) was to design an assistive tool for clinicians to help them spend less time diagnosing difficult-to-detect*PCOS*. However, the selection criteria of the classification models is obfuscated . In the same vein, Cüvitoglu and Isik in (2018) evaluated several classification schemes for cryotherapy and immunotherapy datasets that are applied to wart treatment. The unbalanced classes were balanced by oversampling the class with fewer samples. This work evaluated the pros and cons of all the classification schemes. The Random Forest classifier could reach 0.95, 88, and 98% in accuracy, sensitivity, and specificity, respectively. In future versions of their work, the authors hope to achieve better results with a multiclass One-Class (OC) classifier that essentially targets unbalanced datasets.

Out of the two prominent approaches to detecting polycystic ovary (PCO) follicles, namely stereological (3D objects' projection in 2D planes as images) analysis and crafted feature recognition, Wisesty and Nasri in (2016) utilized the latter approach. Here, the follicle orthogonality is captured by the Gabor wavelet and classified by the back-propagation neural network (BPNN). The BPNN is fine-tuned using Levenberg-Marquardt optimization and the Conjugate Gradient. The Levenberg-Marquardt optimization (accurate by 94%) outclasses

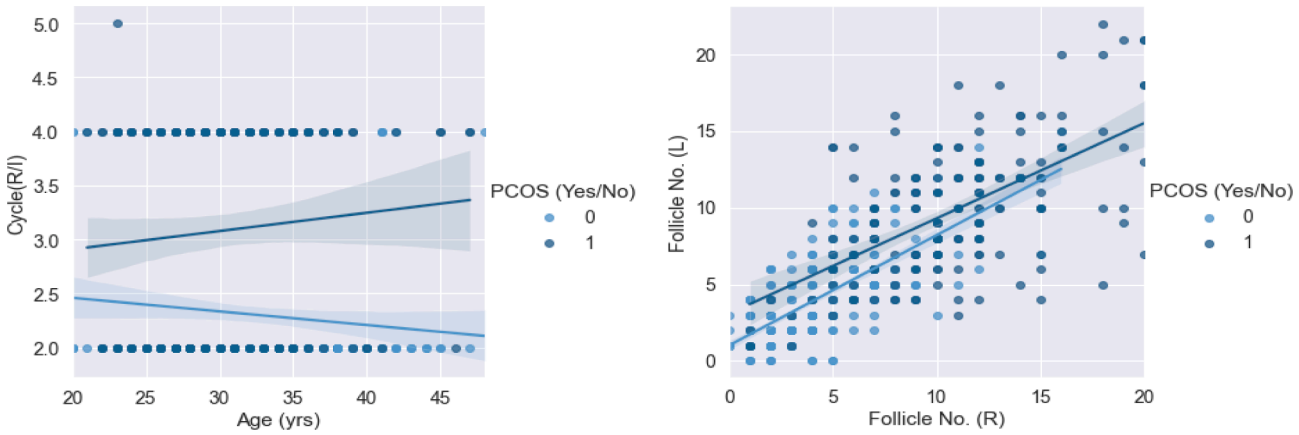

Fig. 4. (a) Visualizes the linear relationship as determined through regression for cycle IR concerning age, (b) Visualizes the linear relationship as determined through regression for distribution of follicles in both ovaries.

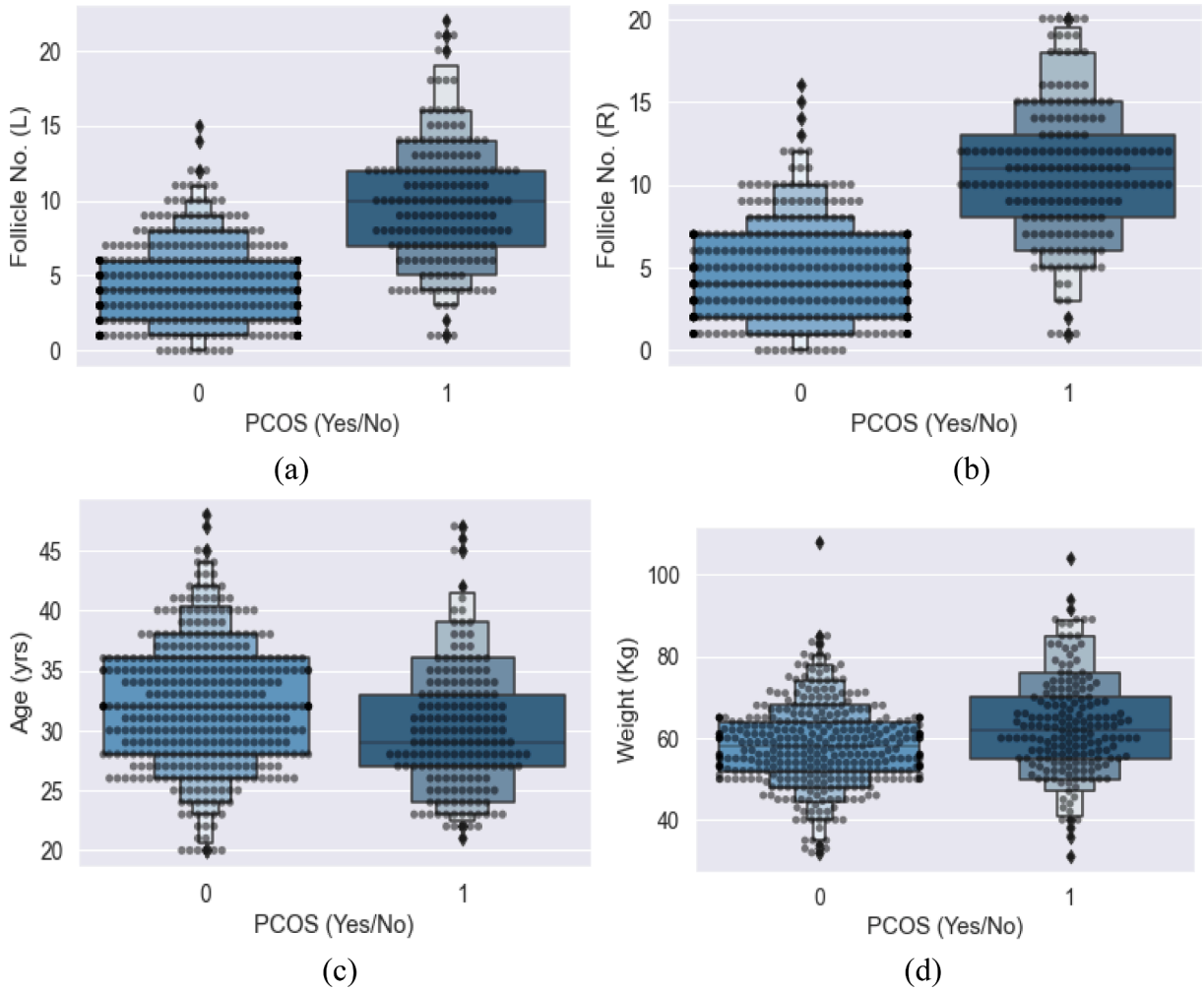

Fig. 5. Swarm plots to display the joint distribution of (a) Follicle No. (L), (b) Follicle No. (R), (c) Age (yrs), (d) Weight (Kg), (e) BMI, (f) Hb(g/dl), (g) Cycle length (days), (h) Endometrium (mm) and target variable PCOS (Yes/No).

Conjugate Gradient-based optimization of BPNN.

PCO ovaries are composed of several follicles or tiny cysts. Another *PCOS* classification work based on follicle detection by Purnama et al. in (2015) obtains binary follicle images using significant low-level image preprocessing by filters and histograms. The segmented follicle blobs are processed for general moment-based features, for example mean, variance, kurtosis, skewness, etc. Classification based on these features leads to the differentiation between *PCOS* and *Non* −*PCOS* follicles. A neural network with learning vector quantization, a Euclidean distance-based k nearest neighbors, and an SVM with radial basis function (RBF) kernel

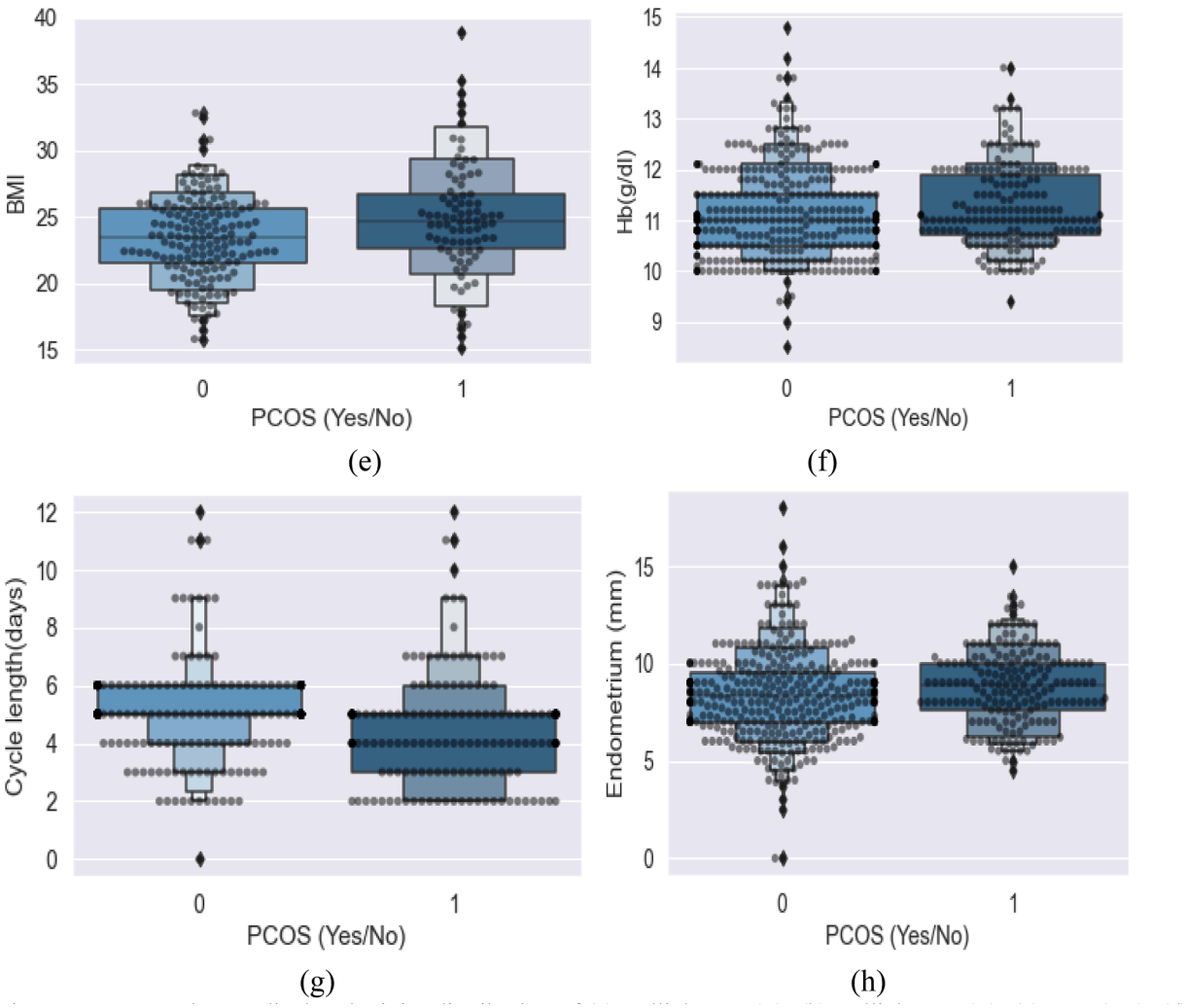

Fig. 5. (*continued*).

were used as classifiers. Experiments were conducted on two datasets of 239 and 339 follicle images each. The best accuracies achieved on the two datasets were 82.5 and 78.8%.

Setiawati and Tjokorda in (2015) propose follicle segmentation using Particle Swarm Optimization (PSO) with an improved non-parametric fitness function to generate more convergent clusters. Enhanced convergence is experimentally justified by comparing it to a non-parametric fitness function. The effectiveness was evident on ultrasound images. Contrast enhancement was shown to improve PSO image clustering with a larger intra-cluster distance. The scheme produced a closer Region of Interest (ROI) to the reference ROI marked by the doctor. Further, classification was carried out on unknown ultrasound ovarian images using logistic regression, SVM, and BPNN (Setiawati et al., 2016). It was noted that BPNN gave the best performance (F-measure 0.85).

The work by Denny et al. in (2019) targeted prompt recognition of *PCOS* from clinical and metabolic parameters. The dataset was collected from a survey conducted during consultations and medical check-ups of 541 female patients. The selected feature-set, transformed using Principal Component Analysis (PCA), was classified using Naïve Bayes, logistic regression, KNN, Random Forest, Classification and Regression Trees (CART), and SVM classifiers. The RF classifier performed the best, with an accuracy of 89.02%.

Satish et al. in (2020) supported the assertion that the RF classifier outperforms other models (claiming 93% accurate results). The classification was performed using the Python-Scikit Learn package and the RapidMiner tools. KNN and SVM perform nearly as well, with around 91% accuracy. Meena et al. in (2015) presented a Neural Fuzzy Rough Sets (NFRS) and ANN-based method that helped avoid the separate classification and feature selection tasks. It was asserted that compared to the established classifiers like SVM, Naïve Bayes, and classification trees, fuzzy ANN schemes perform better. It was shown that NFRS attributes were fewer compared to information gain, gain ratio methods, correlation-based feature selection, and PCA. The presented method was performant for early *PCOS* detection.

The principal aim of the work by Silva et al. in (2021) was to select the most prominent clinical and laboratory variables for *PCOS* diagnosis. The feature selection and the prediction were done by the BorutaShap method and the RF classifier. 58 selected features were ranked as per their importance and claimed accuracy of 86%. Conventionally, the follicle count is manually determined and therefore biased by human perception. Overlapping of follicles in the ultrasonographic examination may also cause misdiagnosis. The work by Padmapriya and Kesavamurthy in (2015) presented a method of automatic detection and follicle counting in the ovarian ultrasound image. This is an image processing-based approach for preprocessing and region of interest extraction that determines the size and number of follicles in ultrasound images. Image processing results classify whether an ovary is *PCOS* affected. Thus, the method avoids human-based manual tracing of follicles, resulting in increased efficiency and accuracy of the system.

Work by Zhang et al. in (2021) identifies samples affected by *PCOS* by performing Raman spectroscopy with PCA and spectral classification

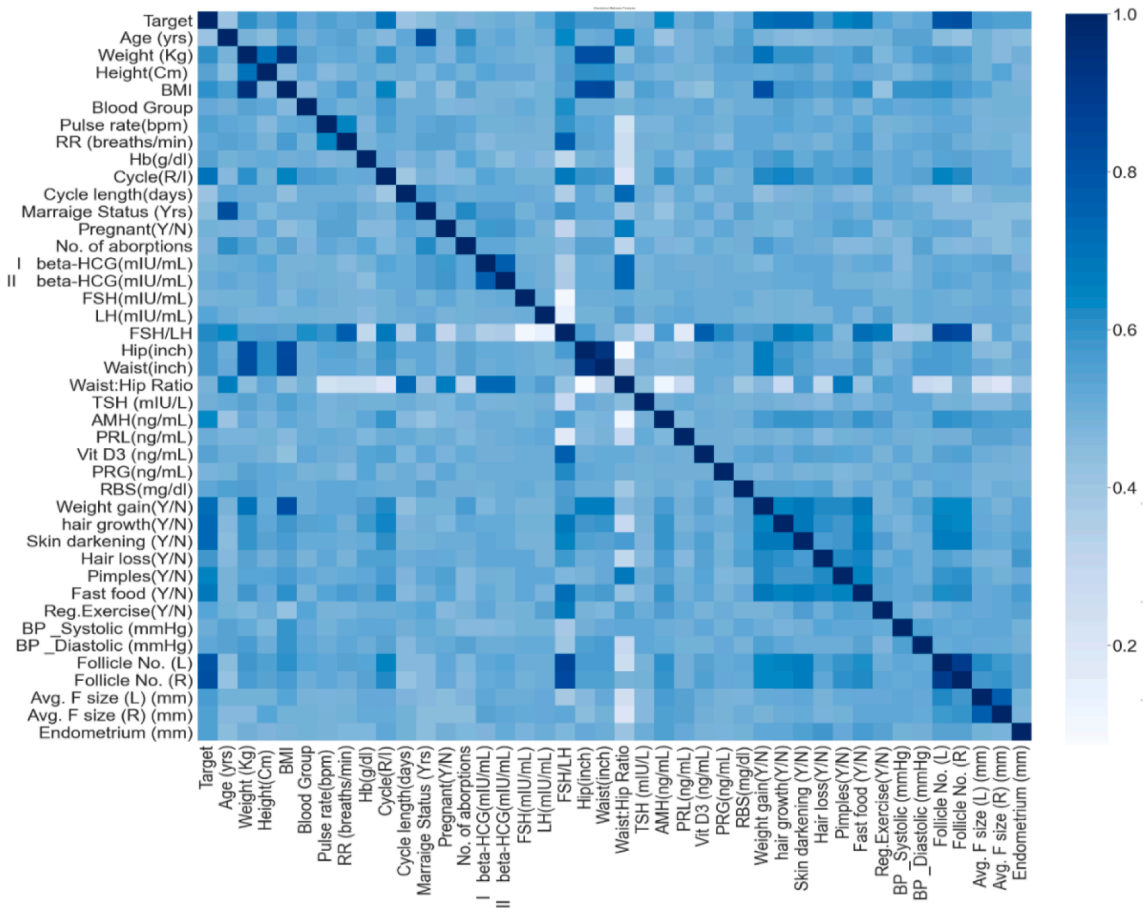

Fig. 6. Heatmap for examining the correlations between all the features of the Polycystic Ovary Syndrome dataset.

models. Follicular fluid and plasma samples from 50 women (*PCOS* and *Non −PCOS*) were subjected to Raman spectroscopy. Discrimination between the *PCOS* and *Non −PCOS* groups was evident through the PCA analysis. A stacking classification model based on RF, extreme gradient boosting, and KNN resulted in 89% accuracy on follicular liquid, while 75% accuracy was obtained on plasma samples. Conclusively, Raman spectroscopy is a unique and cost-effective technique for automated *PCOS* diagnosis. Katarya et al. in (2021) developed an automated *PCOS* prognosis system where feature selection was carried out using PSO and a customized stacked ensemble-learning model. This system was 90.74% accurate.

The rigorous navigation through the pertinent literature reflects the need for further evaluation and cross-examination. Therefore, the proposed work carries out a rigorous evaluation with a set of established and cutting-edge classifiers.

3. Material and methods

This section discusses the designs of classifiers utilized for *PCOS* classification, data set, and exploratory data analysis.

3.1. Methodology

The complete methodology for *PCOS* diagnosis is depicted in Fig. 1. After data cleaning and feature selection, a collection of classifiers such as Support Vector Machine (SVM) with linear and radial kernels, Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Gradient Boosting (GB), AdaBoost (AB), XGBoost (XB) and CatBoost (CB) are analyzed for PCOS classification. These models are tested for a variable level of correlation coefficients. After model evaluation, the best model and optimum correlation coefficient are selected. Finally, that best model is tuned to achieve the final results.

3.2. Machine learning classifiers

A brief overview of the analyzed classifiers utilized for *PCOS* classification is discussed below.

- **Support Vector Machine:** SVM is a supervised machine learning algorithm that can be used for classification and regression (Tiwari et al., 2021). SVM's divides datasets into classes so that a maximumally marginal hyperplane can be identified. It transforms data by using the kernel method and then generates an ideal boundary between the potential outputs based on the changes. The kernel is a set of mathematical functions that are used in SVM approaches. Different types of kernel functions are used by different SVM algorithms. Various types of functions which can be used are linear, nonlinear, polynomial, sigmoid, and Radial Basis Function (RBF). Two types of kernels are used in this work, i.e., linear and radial kernels. The kernel function is given in equation (1), here d is the degree of the polynomial.

$$K(x,y) = (1 + \sum_{j=1}^{p} x_{i,j} y_{i,j})^{d} \quad (1)$$

The Radial kernel is given in equation (2), here γ is a tuning parameter that decides the smoothness of the decision boundary and controls the variance of the model.

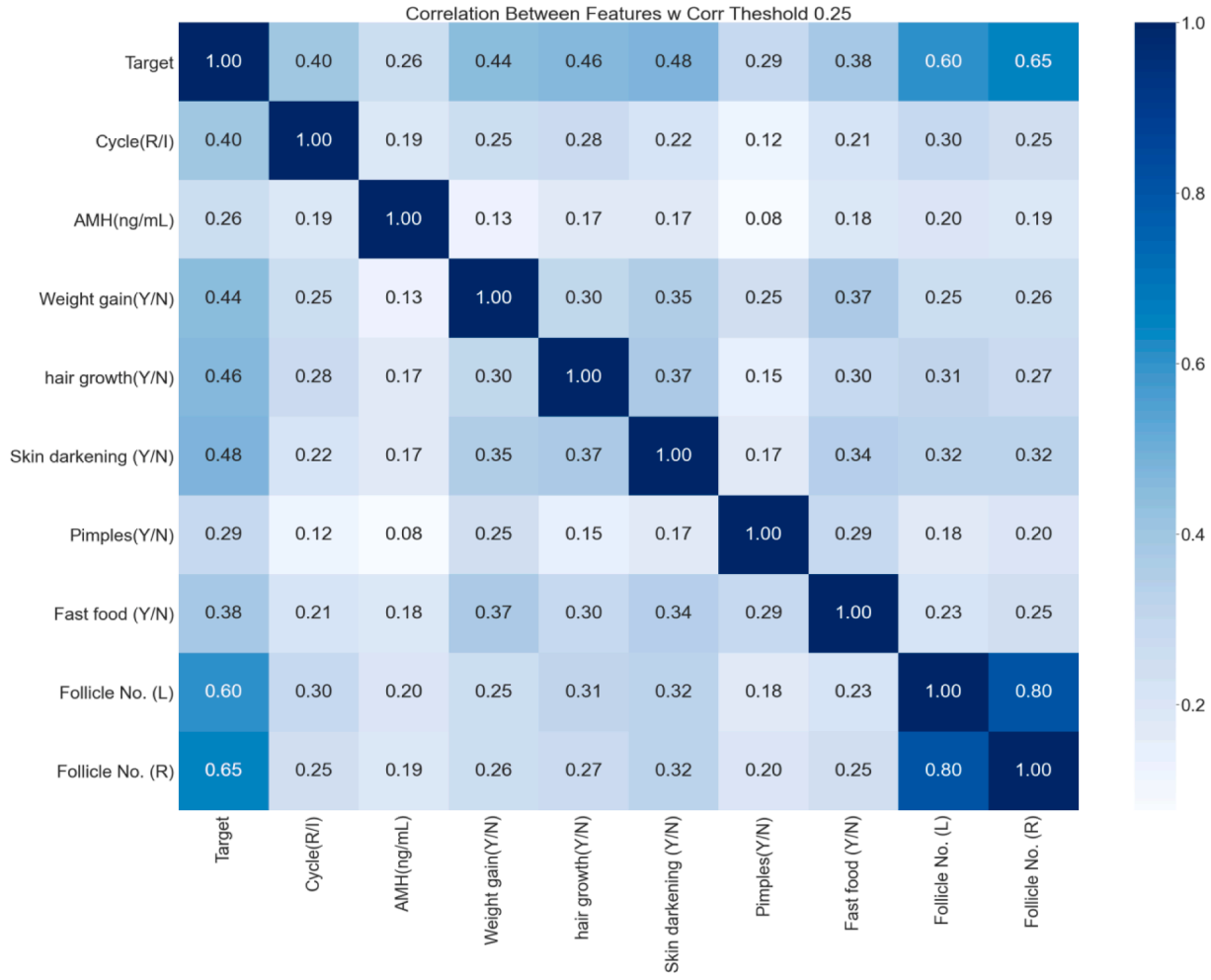

Fig. 7. Heatmap for the correlations between the features with threshold 0.25 of Polycystic Ovary Syndrome dataset.

Table 1
Performance of various classifiers in terms of accuracy at different threshold levels of correlation for Polycystic Ovary Syndrome detection.

Classifier/Threshold Level	0.1	0.2	0.3	0.4	0.5	0.6	0.8	0.9	0.95
Linear SVM	0.723	0.723	0.754	0.84	0.858	0.871	0.871	0.871	0.846
Radial SVM	0.711	0.711	0.711	0.711	0.711	0.711	0.711	0.711	0.711
Logistic Regression	0.699	0.736	0.693	0.803	0.809	0.815	0.858	0.858	0.871
Random Forest	0.723	0.742	0.76	0.877	0.895	0.901	0.924	0.895	0.907
AdaBoost	0.711	0.674	0.687	0.809	0.815	0.828	0.858	0.858	0.858
Decision Tree	0.619	0.644	0.693	0.852	0.852	0.852	0.871	0.858	0.858
K-Nearest Neighbors	0.668	0.65	0.644	0.662	0.668	0.668	0.674	0.674	0.687
Gradient Boosting	0.699	0.693	0.705	0.858	0.883	0.865	0.877	0.889	0.889
Xgboost	0.717	0.699	0.742	0.865	0.865	0.858	0.877	0.877	0.883
CatBoost	0.754	0.711	0.736	0.842	0.877	0.865	0.907	0.901	0.883
Linear Discriminant Analysis	0.736	0.723	0.742	0.858	0.883	0.889	0.907	0.907	0.907
Quadratic Discriminant Analysis	0.682	0.717	0.723	0.815	0.822	0.834	0.865	0.84	0.852

$$K(x,y) = exp(-\gamma \sum_{j=1}^{p} (x_{i,j} - y_{i,j})^2) \quad (2)$$

- **Decision Tree:** A decision tree is a supervised learning classifier that can be used for both regression and classification. In this tree-structured classifier (Elmachtoub et al., 2020), internal nodes reflect dataset features, branches reflect decision rules, and every leaf node delivers the conclusion. The classification decision tree is built using binary recursive partitioning. Dividing the data into partitions and then further breaking it up on each branch is part of this iterative process.
- **Random Forest Classifier** is a regression and classification algorithm that is an ensemble of decision tree algorithms. It utilizes bagging and feature randomization to produce an uncorrelated forest of trees whose aggregate prediction is more exact than the forecast of any one tree. It makes use of ensemble learning, a technique for resolving complex issues through the employment of several classifiers (Roy et al., 2020).
- **Logistic regression** is a supervised classification approach used to predict the likelihood of a target variable. Because the fundamental technique is quite similar to linear regression, it is called "logistic regression." The word "logistic" refers to the Logit function, which is employed in this categorization method. Because the nature of the

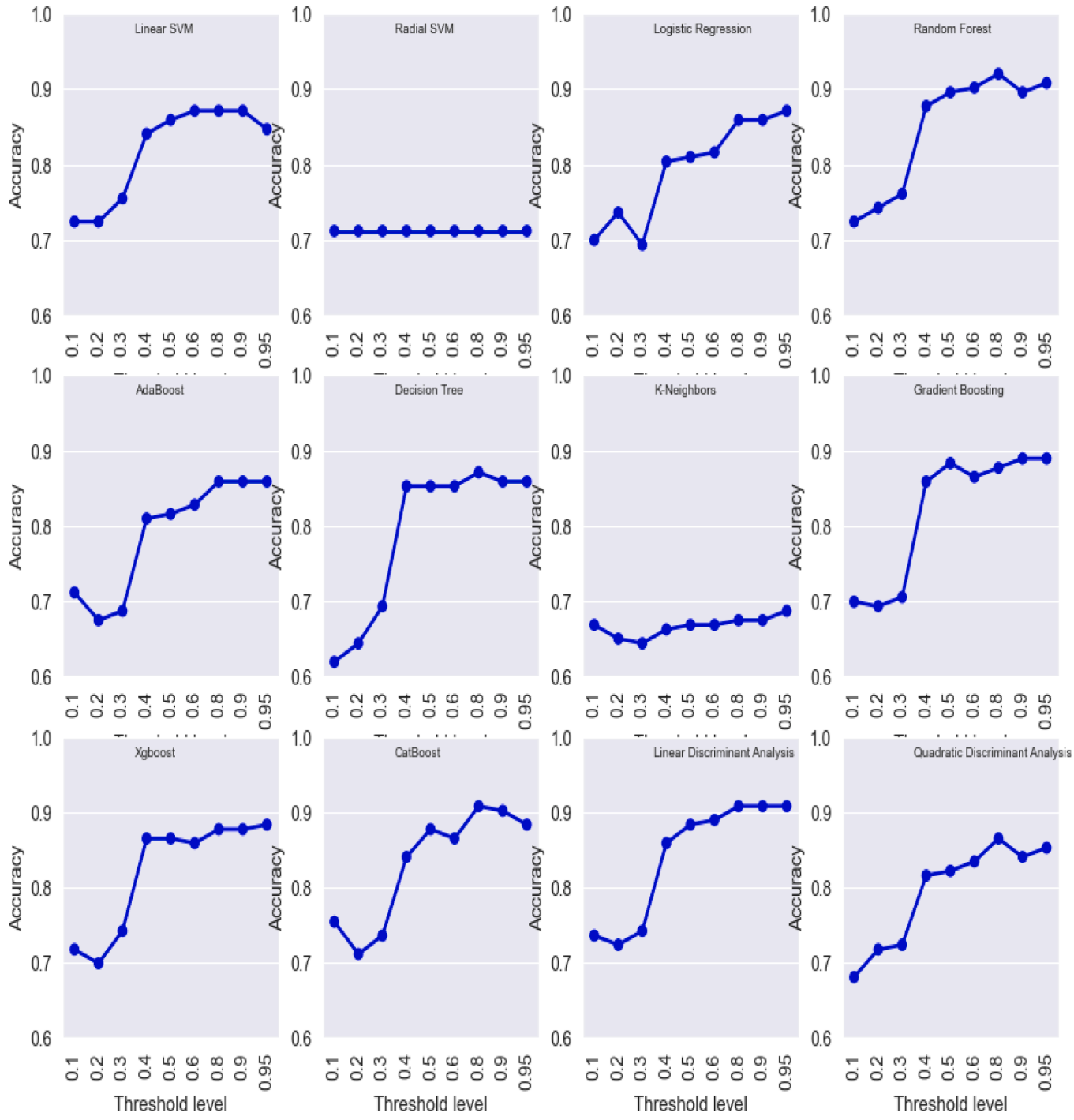

Fig. 8. Accuracy plots of various classifiers at different threshold levels of correlation for Polycystic Ovary Syndrome detection.

goal or dependent variable is dichotomous, there are only two classes (Nusinovici et al., 2020).

- **K-Nearest Neighbors:** The K-Nearest Neighbor classifier is a supervised machine learning method. Both regression and classification problem statements can be solved using the approach. Computing the lengths among a query and all of the examples in the data, choosing the nearest neighbor instances to the query, and voting for the most common label for classification are all part of the KNN algorithm. Minkowski distance is used for distance calculations and it can be defined as in Eq. (3) where p is the order of the norm (Shah et al., 2020).

$$d = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} \quad (3)$$

- **Linear Discriminant Analysis:** Discriminant analysis is a term that refers to a collection of methods for dimensionality reduction and classification. In statistical and probabilistic learning, linear discriminant analysis and quadratic discriminant analysis are two well-known supervised classification approaches. LDA is a method that uses a linear combination of features to distinguish two or more groups of samples (Ricciardi et al., 2020). The linear designation is due to the linearity of the discriminant functions. LDA does

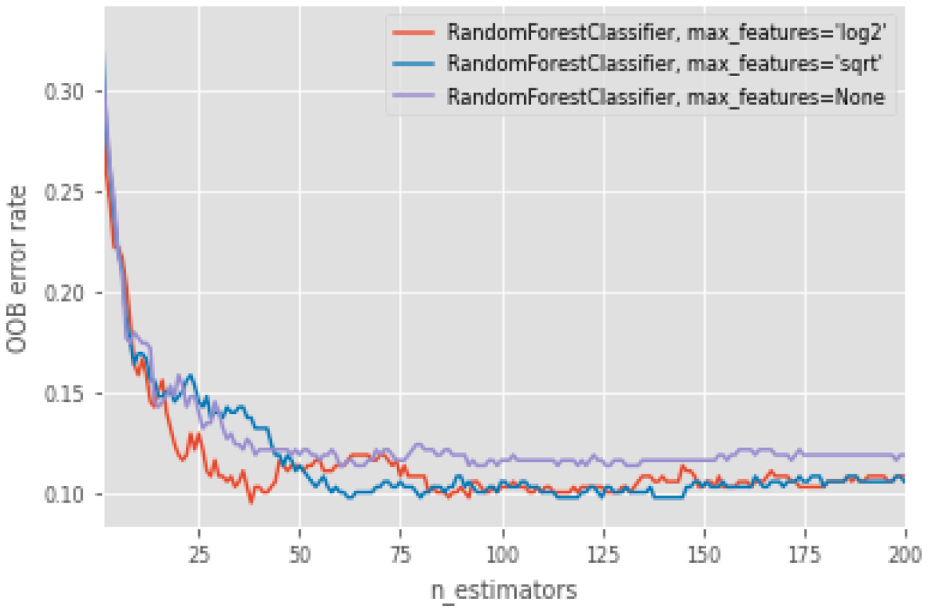

Fig. 9. Exhibits how the OOB error is estimated at the addition of each new tree during training with different feature bagging strategies. The error stabilizes at 140 number of estimators for these feature bagging strategies.

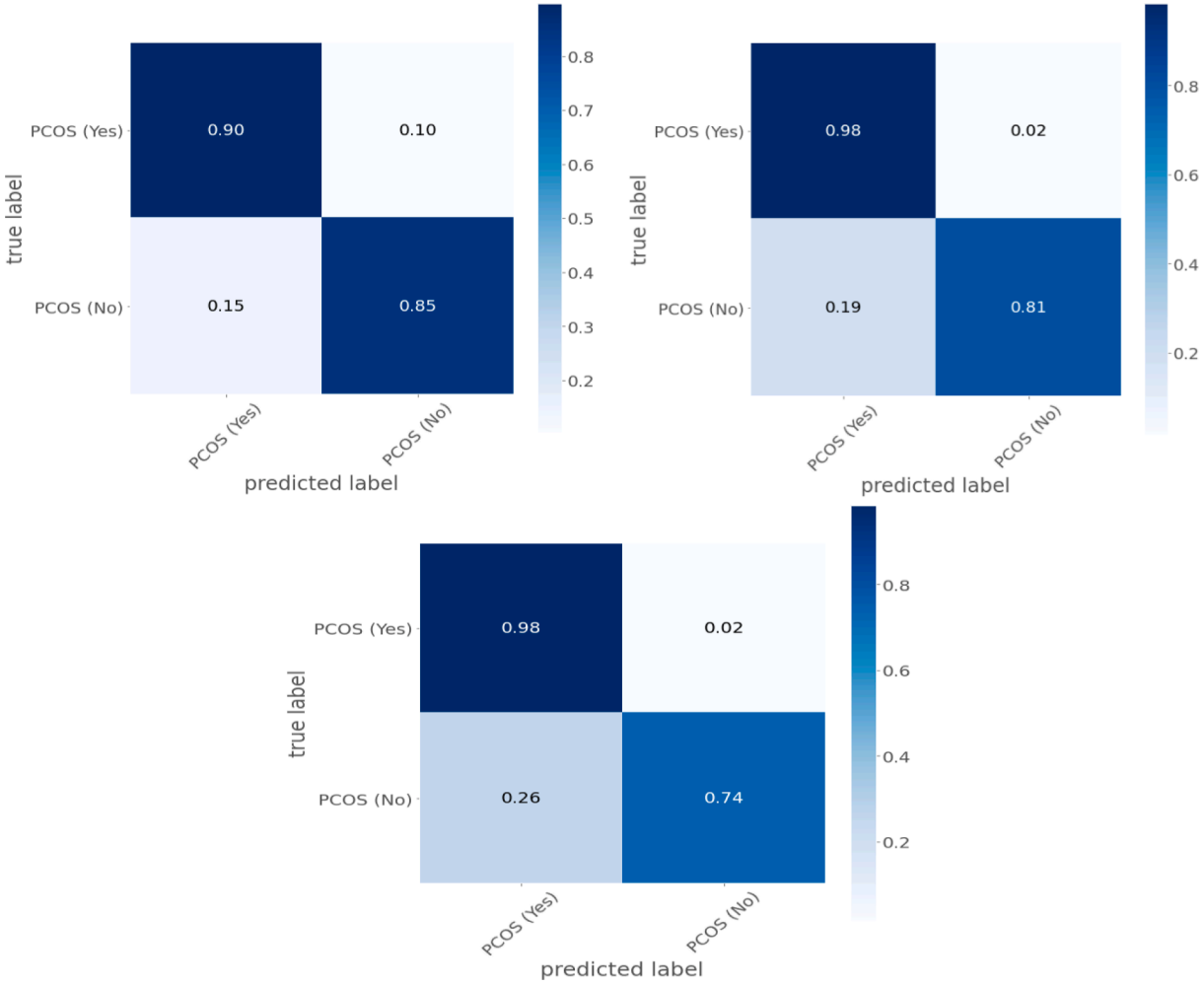

Fig. 10. Heatmap plots of confusion matrices for (a) Random Forest classifier with None as maximum features hypermeter, (b) Random Forest classifier with sqrt as maximum features hypermeter, (c) Random Forest classifier with log2 as maximum features hypermeter.

Table 2
Results for Random Forest classifier with None, sqrt, and log2 as maximum features hypermeter for Polycystic Ovary Syndrome classification.

Class/Metric	Random Forest classifier with None as maximum features hypermeter			Random Forest classifier with sqrt as maximum features hypermeter			Random Forest classifier with $\log_2$ as maximum features hypermeter		
	Precision	*Recall*	*F −Score*	*Precision*	*Recall*	*F −Score*	*Precision*	*Recall*	*F −Score*
PCOS without Infertility	89.66	94.55	92.04	98.28	92.68	95.40	98.28	90.48	94.21
PCOS with Infertility	87.23	77.36	82.00	80.85	95.00	87.36	74.47	94.59	83.33
macro avg. accuracy	88.44	85.95	87.02	89.56	93.84	91.38	86.37	92.54	88.77
weighted avg. accuracy	88.87	88.96	88.77	94.00	93.25	93.42	92.87	91.41	91.74
Accuracy	88.96			93.25			91.41		

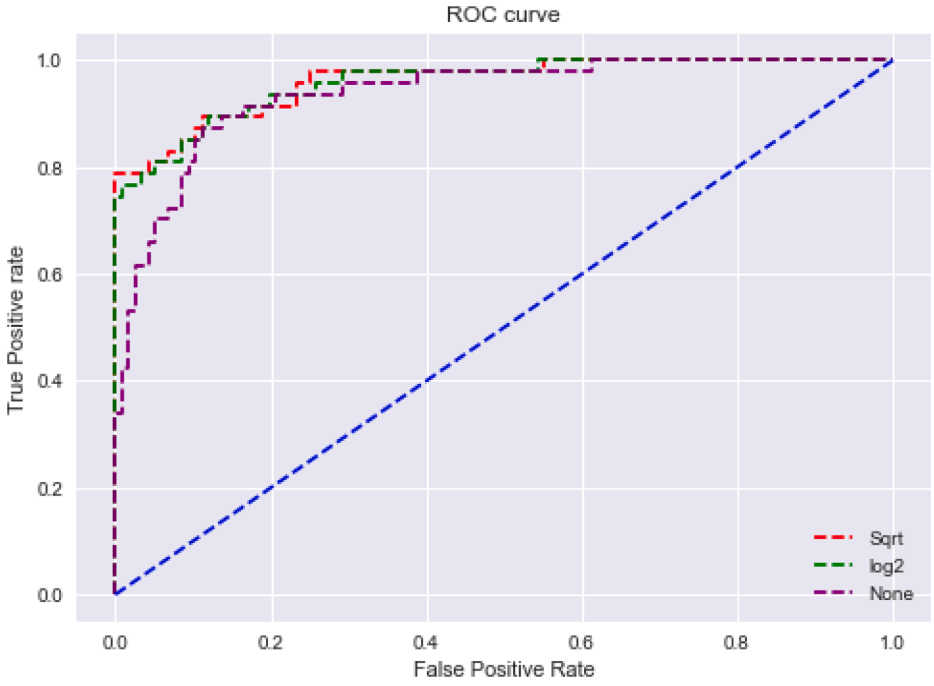

Fig. 11. ROC plots for different feature bagging strategies for hypermeter for Polycystic Ovary Syndrome classification.

this with a three-step procedure that focuses on projecting features from a higher-dimensional space to a lower-dimensional space.

- **Quadratic Discriminant Analysis:** Quadratic Discriminant Analysis is a generative model. Each class in QDA is assumed to have a Gaussian distribution. The proportion of data points that belong to the class is the class-specific prior. The average of the class-specific input variables makes up the class-specific mean vector. When a linear border between classifiers is necessary, LDA is employed, whereas QDA is used when a non-linear boundary is required (Wong et al., 2020).
- **Gradient Boosting Classifier:** Gradient boosting classifiers are a group of machine learning procedures that merge numerous weak learning models to generate a powerful predictive model. Decision trees are commonly employed for gradient boosting. Gradient boosting is a technique for sequential training of several models. It's a variant of boosting that applies to any differentiable loss function. It can be applied to solve problems involving regression and classification (Bahad & Saxena, 2020).
- **AdaBoost Classifier:** The AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. These are models that achieve accuracy just above random chance when applied to classification issues. One-level decision trees are the most suitable and thus the most frequently utilised strategy with AdaBoost. It's a form of *meta*-estimator that starts by fitting a predictor to the given dataset, then adapts successive copies of the predictor to almost the same dataset while updating the weights of incorrectly classified instances to guarantee that following classifiers emphasis on challenging complex scenarios (Rehman Javed et al., 2020).
- **XGBoost Classifier:** XGBoost acronym for extreme Gradient Boosting is a scalable and accurate gradient boosting system implemented using a decision-tree-based ensemble Machine Learning method. This approach works well for a variety of classification and regression predictive modeling issues. It's quick employment of the stochastic gradient boosting procedure with several hyperparameters for model training in a fine-grained control manner (Yu et al., 2020).
- **CatBoost Classifier:** CatBoost is a decision tree gradient boosting technique. The term CatBoost is an acronym that stands for 'Category' and 'Boosting'. Gradient boosting learning is referred to as "boosting" in CatBoost. Gradient boosting is a classification and regression machine learning approach. A series of decision trees are constructed one after the other during training. Each succeeding tree is constructed with less loss than the prior trees. The initial settings determine the number of trees. It has the potential to improve model performance while lowering overfitting and tweaking time. This approach is fast and implements a greedy new gradient boosting algorithm (Ibrahim et al., 2020).

3.3. PCOS dataset

The diagnosis data for Polycystic Ovary Syndrome used in this work is a subset of the dataset supplied by Kottarathil and accessible via the

Kaggle repository (Kottarathil, 2020). This data set was compiled from ten distinct institutions in Kerala, India, and contains diagnostic and clinical information on 541 individuals. The dataset is divided into 70:30 training and testing portions. The file contains the physical and clinical *PCOS* factors used in this work, as it is primarily concerned with screening and diagnosis. This data contains 42 parameters, two of which have been identified as unique identifier values and have been removed. Marriage status was deemed unnecessary for the remaining 40. The target variable "*PCOS* " is binary, with one representing a positive case and zero representing a non-positive case. The data was cleaned by eliminating records from a patient with inconsistent, non-numeric data, resulting in a final dataset of 540 items. To maintain a balanced distribution of classes in the testing and training datasets, we used the Synthetic Minority Over-sampling Technique (SMOTE) to resample our data. We then split the balanced dataset at a 70:30 ratio for training and testing. The following section discusses exploratory data analysis and correlation analysis.

3.4. Exploratory data analysis and feature selection

The remaining 38 features are screening and diagnostic criteria. Screening parameters are a collection of non-invasive diagnostic tests that do not require the patient's sample to be extracted. Among the screening parameters are the patient's history, physiological, and metabolic data. These consist of: age, weight, height, body mass index, blood group, pulse rate, respiratory rate, cycle regularity, cycle length, pregnancy, number of abortions, hip, waist, waist-to-hip ratio, weight gain, skin darkening, hair loss, pimples, consumption of fast food, regular exercise, and blood pressure systolic and blood pressure diastolic. A diagnostic parameter, on the other hand, is a follow-up test that may involve a fluid sample or an invasive vaginal ultrasound. Hb, FSH, LH, FSH/LH, TSH, AMH, PRL, Vitamin D3, PRG, RBS, Follicle No. (L), Follicle No. (R), Avg. F size (L), Avg. F size (R), and Endometrium are all diagnostic characteristics. Figs. 2–5 illustrate various plot types used in exploratory data analysis, including frequency distribution plots, regression plots, and swarm plots.

When building a classifier, the constraint of dimensionality must be properly addressed. One significant challenge is selecting the best collection of features for classification given a high -dimensional dataset. We identified features based on their correlation, or how closely two variables vary together. A correlation heatmap between all the features of the *PCOS* dataset is provided in Fig. 6. Each cell in this heatmap corresponds to the correlation coefficient between two variables. Pearson's coefficient (ρ), a measure or linear correlation, is computed for correlation analysis. It is computed by dividing the covariance between two variables by the product of standard deviations (σ_x, σ_y) of the two variables x and y as defined in Eq. (4) (Wang et al., 2020).

$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{4}$$

Here, $cov(x,y)$ is covariance between x and y and defined as the 'expected value of the product of the deviations of x and y from their respective means (μ_x, μ_y). in Eqs. (5) and (6).

$$cov(x,y) = E\left[(x-\mu_x)(y-\mu_y)\right] \tag{5}$$

So, in other terms.

$$\rho(x,y) = \frac{E\left[(x-\mu_x)(y-\mu_y)\right]}{\sigma_x \sigma_y} \tag{6}$$

We examine feature correlations and eliminated one of two features with a correlation greater than certain thresholds. The values for the thresholds range from 0.1 to 0.95. Fig. 7 provides the heatmap for the correlations between the features with a threshold of 0.25.

4. Experiment and results

The machine learning classifiers, namely SVM, DT, RF, LR, GB, AB, XB, and CB, are analyzed for *PCOS* classification for different levels of correlation coefficients. The performance of these models is compared based on accuracy, as recorded in Table 1. The plots of accuracy are provided in Fig. 8. Observations deduced from these results are discussed below.

- The Random Forest classifier achieved the highest performance, with a 92.4% accuracy at a correlation threshold of 0.8.
- Linear SVM classifiers achieved a maximum accuracy of 87.1% at correlation thresholds 0.6, 0.7, and 0.8.
- The radial SVM classifier maintains a constant accuracy of 71.1% with varying correlation levels.
- The K-Nearest Neighbors classifier maintains a constant accuracy of approximately 66% with varying correlation levels.
- The Linear Discriminant Analysis and CatBoost classifiers also achieved noteworthy performance with accuracies greater than 90%.

As discussed in Section 4, the Random Forest classifier outperformed all other models with its correlation threshold set to 0.8. The RF classifier is further tuned to optimize for low out-of-bag (OOB) error. During the training phase, the OOB error is used to approximate the random forest error (Li et al., 2021). A different bootstrap sample is used for each tree. Approximately one-third of the observations are randomly excluded from each bootstrap sample. The OOB sample refers to the observations that were left out of a certain tree. The hyperparameters are fine-tuned to identify the RF classifier with the best testing precision. Also fine-tuned are the number of repetitions (total subtrees) and other factors to explore randomly at individually splitting because out-of-bag error rates converge as the number of repetitions approaches infinity. It is only necessary to set the repetitions to a value adequate for convergence. n_estimators, the number of trees used in the forest, is adjusted to control the number of trees to be engaged in the procedure. We run the random forest function iteratively to show how the out-of-bag error and validation error have similar patterns as the number of repetitions increases. The variable number of iterations is set to one and grows by one per function call until it reaches 200 for different feature bagging strategies, namely log base 2, square root, and None. Finally, the values of out-of-bag error and validation error alongside the number of repetitions are displayed, as in Fig. 9, to visualize the trends. When splitting a node without replacement, max features is the size of the random subsets of features to evaluate. The plot in Fig. 9 verifies that square root feature bagging outperforms other schemes, and the errors stabilize after 140 iterations.

Further, a Random Forest classifier with all three feature bagging schemes and 140 trees is assessed on precision (P), sensitivity (S), F-score (F), and accuracy (A). These performance metrics are discussed below in Eqs. (7)–(10) (Tiwari et al., 2021; Tiwari, 2021).

$$P = T_{PCOS}/(T_{PCOS} + F_{PCOS}) \tag{7}$$

$$S = T_{PCOS}/(T_{PCOS} + F_{Non-PCOS}) \tag{8}$$

$$F = (2{*}P{*}S)/(P+S) \tag{9}$$

$$A = (T_{PCOS} + T_{Non-PCOS})/(T_{PCOS} + T_{Non-PCOS} + F_{PCOS} + F_{Non-PCOS}) \tag{10}$$

Where T_{PCOS} are the samples when somebody is *PCOS*-positive, and *PCOS*, F_{PCOS} are the samples when somebody is *PCOS*-negative. *PCOS*, $T_{Non-PCOS}$ are the *Non −PCOS* sample misclassified as *Non −PCOS*, and $F_{Non-PCOS}$ are the *PCOS* samples misclassified as *Non −PCOS*.

To evaluate these metrics, the first confusion matrix is computed in Fig. 10. In a classification problem, a confusion matrix describes the prediction values vs. the actual values. Correct and erroneous

predictions are listed in a table with their respective values, divided down by class. Table 2 provides the results based on these performance metrics. These metrics show that a Random Forest classifier with 'sqrt' as maximum features hypermeter outperforms other feature bagging strategies. It offers 93.25% accuracy which is better than accuracies provided Random Forest classifier with None and *log*2 as maximum features hypermeter – 88.96 and 91.41% respectively.

The results are further confirmed using the area under the ROC curve (AUC). The area under the ROC curve is used as an holistic measurement to evaluate the classification performance, and it is generated by plotting along a series of decision thresholds. While standard accuracy is determined by a single cut point, ROC tests all of the decision thresholds and depicts the sensitivity and specificity (Janssens & Martens, 2020). The plots in Fig. 11 show that the square root feature bagging technique slightly outperforms other bagging schemes.

5. Conclusion

Polycystic Ovarian Syndrome is a complicated health problem with too complex a space of symptoms to easily diagnose clinically. Proper diagnosis is the foundation of an effective treatment, and in this work, we designed experiments using machine learning for efficiently diagnosing PCOS based on patient clinical data. We identified statistically significant and discriminating features based on the correlation coefficient that best indicated PCOS and then fed the features into different machine learning models. According to our extensive experiments on a benchmark dataset, the Random Forest had the greatest power in PCOS diagnosis. We reached this conclusion by using out-of-bag error to evaluate the accuracy of different models and selecting optimal values for tuning parameters, such as the number of representative predictors picked at random for a split. In Random Forest classification, the out-of-bag error was measured to choose tuning parameters such as tree splitting strategy and number of trees. Despite the success in this study, and because performance is dependent on the nature of the dataset, sampling, and pre-processing procedures, Random Forests may not always be the best solution. This study's future scope could include the use of multi-modality data sets for PCOS diagnosis, such as ultrasound scans, as well as the use of different or larger data sets for diagnosis.

CRediT authorship contribution statement

Shamik Tiwari: Conceptualization, Methodology. **Lalit Kane:** Visualization. **Deepika Koundal:** Data curation, Investigation. **Anurag Jain:** Writing – original draft. **Adi Alhudhaif:** Software. **Kemal Polat:** Writing – review & editing. **Atef Zaguia:** Validation. **Fayadh Alenezi:** Supervision. **Sara A. Althubiti:** Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported by Taif University Researchers Supporting Project Number (TURSP-2020/114), Taif University, Taif, Saudi Arabia.

References

Azziz, R., Carmina, E., Chen, Z., Dunaif, A., Laven, J. S., Legro, R. S., … Yildiz, B. O. (2016). Polycystic ovary syndrome. *NatureReviews Disease Primers, 2*(1), 1–18.

Bahad, P., & Saxena, P. (2020). In *Study of adaboost and gradient boosting algorithms for predictive analytics* (pp. 235–244). Singapore: Springer.

Barber, T. M., & Franks, S. (2021). Obesity and polycystic ovary syndrome. *Clinical Endocrinology, 95*(4), 531–541.

Bharati, S., Podder, P., & Mondal, M. R. H. (2020 June). In *Diagnosis of polycystic ovary syndrome using machine learning algorithms* (pp. 1486–1489). IEEE.

Bharati, S., Podder, P., Mondal, M., Surya Prasath, V. B., & Gandhi, N. (2022). In *Ensemble Learning for Data-Driven Diagnosis of Polycystic Ovary Syndrome* (pp. 1250–1259). Cham: Springer.

Bhardwaj, P., & Tiwari, P. (2022). In *Manoeuvre of Machine Learning Algorithms in Healthcare Sector with Application to Polycystic Ovarian Syndrome Diagnosis* (pp. 71–84). Singapore: Springer.

Cüvitoglu, A., & Isik, Z. (2018). Evaluation machine learning approaches for classification of cryotherapy and immunotherapy datasets. *International Journal of Machine Learning and Computing, 4*(4), 331–335.

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

Denny, A., Raj, A., Ashok, A., Ram, C. M., & George, R. (2019 October). In *I-HOPE: detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques* (pp. 673–678). IEEE.

Elmachtoub, A., Liang, J. C. N., & McNellis, R. (2020, November). Decision trees for decision-making under the predict-then-optimize framework. In *International Conference on Machine Learning* (pp. 2858-2867). PMLR.

Escobar-Morreale, H. F. (2018). Polycystic ovary syndrome: Definition, aetiology, diagnosis and treatment. *Nature Reviews Endocrinology, 14*(5), 270–284.

Guleken, Z., Bulut, H., Bulut, B., Paja, W., Orzechowska, B., Parlinska-Wojtan, M., & Depciuch, J. (2022). Identification of polycystic ovary syndrome from blood serum using hormone levels via Raman spectroscopy and multivariate analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 273*, 1–10. https://doi.org/10.1016/j.saa.2022.121029

Hart, R., & Doherty, D. A. (2015). The potential implications of a PCOS diagnosis on a woman's long-term health using data linkage. *The Journal of Clinical Endocrinology & Metabolism, 100*(3), 911–919.

Hassan, M. M., & Mirza, T. (2020). Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome. *International Journal of Computer Applications, 975*, 42–53.

Ibrahim, A. A., Ridwan, R. L., Muhammed, M. M., Abdulaziz, R. O., & Saheed, G. A. (2020). Comparison of the CatBoost Classifier with other Machine Learning Methods. *International Journal of Advanced Computer Science and Applications, 11*(11).

Janssens, A. C. J., & Martens, F. K. (2020). Reflection on modern methods: Revisiting the area under the ROC curve. *International Journal of Epidemiology, 49*(4), 1397–1403.

Jaralba, J. R., Baldovino, R., & Co, H. (2020 October). In *A Machine Learning Approach for Initial Screening of Polycystic Ovarian Syndrome (PCOS)* (pp. 517–529). Cham: Springer.

Katarya, R., Jindal, A., Duggal, A., & Shah, A. (2021). In *A Novel Polycystic Ovarian Syndrome Diagnostic System Using Machine Learning* (pp. 555–563). Singapore: Springer.

Kiruthika, V., Sathiya, S., & Ramya, M. M. (2020). Machine learning based ovarian detection in ultrasound images. *International Journal of Advanced Mechatronic Systems, 8*(2–3), 75–85.

Kottarathil, P.: Polycystic ovary syndrome (PCOS). (2020). https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos.

Li, M., Xu, Y., Men, J., Yan, C., Tang, H., Zhang, T., & Li, H. (2021). Hybrid variable selection strategy coupled with random forest (RF) for quantitative analysis of methanol in methanol-gasoline via raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 251*, 119430.

Louwers, Y. V., & Laven, J. S. (2020). Characteristics of polycystic ovary syndrome throughout life. *Therapeutic Advances in Reproductive Health, 14*, 1–9.

Maadi, M., Akbarzadeh Khorshidi, H., & Aickelin, U. (2021). A Review on Human–AI Interaction in Machine Learning and Insights for Medical Applications. *International Journal of Environmental Research and Public Health, 18*(4), 1–27.

Meena, K., Manimekalai, M., & Rethinavalli, S. (2015). Correlation of Artificial Neural Network Classification and NFRS Attribute Filtering Algorithm for PCOS Data. *International Journal of Research in Engineering and Technology, 4*(3), 519–524.

Mehrotra, P., Chatterjee, J., Chakraborty, C., Ghoshdastidar, B., & Ghoshdastidar, S. (2011 December). In *Automated screening of polycystic ovary syndrome using machine learning techniques* (pp. 1–5). IEEE.

Meier, R. K. (2018). Polycystic ovary syndrome. *Nursing Clinics of North America, 53*(3), 407–420.

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., … Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology, 122*, 56–69.

Padmapriya, B., & Kesavamurthy, T. (2015). In *Diagnostic tool for PCOS classification* (pp. 182–185). Cham: Springer.

Purnama, B., Wisesti, U. N., Nhita, F., Gayatri, A., & Mutiah, T. (2015). In *May). A classification of polycystic Ovary Syndrome based on follicle detection of ultrasound images* (pp. 396–401). IEEE.

Rehman Javed, A., Jalil, Z., Atif Moqurrab, S., Abbas, S., & Liu, X. (2020). Ensemble adaboost classifier for accurate and fast detection of botnet attacks in connected vehicles. *Transactions on Emerging Telecommunications Technologies, e4088*.

Ricciardi, C., Valente, A. S., Edmund, K., Cantoni, V., Green, R., Fiorillo, A., … Cesarelli, M. (2020). Linear discriminant analysis and principal component analysis to predict coronary artery disease. *Health Informatics Journal, 26*(3), 2181–2192.

Roy, S. S., Dey, S., & Chatterjee, S. (2020). Autocorrelation aided random forest classifier-based bearing fault detection framework. *IEEE Sensors Journal, 20*(18), 10792–10800.

Satish, C. N., Chew, X., & Khaw, K. W. (2020). Polycystic Ovarian Syndrome (PCOS) classification and feature selection by machine learning techniques.

Setiawati, E., & Tjokorda, A. B. W. (2015 May). In *Particle swarm optimization on follicles segmentation to support PCOS detection* (pp. 369–374). IEEE.
Setiawati, E., Wirayuda, T. A., & Astuti, W. (2016). A Classification of Polycystic Ovary Syndrome Based on Ultrasound Images Using Supervised Learning and Particle Swarm Optimization. *Advanced Science Letters, 22*(8), 1997–2001.
Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research, 5*(1), 1–16.
Shan, B., Cai, J. H., Yang, S. Y., & Li, Z. R. (2015). Risk factors of polycystic ovarian syndrome among Li People. *Asian Pacific Journal of Tropical Medicine, 8*(7), 590–593.
Silva, I. S., Ferreira, C. N., Costa, L. B. X., Sóter, M. O., Carvalho, L. M. L., … Gomes, K. B. (2021). Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models. *Journal of Endocrinological Investigation*, 1–9.
Teede, H. J., Misso, M. L., Costello, M. F., Dokras, A., Laven, J., Moran, L., … Norman, R. J. (2018). Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Human Reproduction, 33*(9), 1602–1618.
Tiwari, S. (2021). Dermatoscopy using multi-layer perceptron, convolution neural network, and capsule network to differentiate malignant melanoma from benign nevus. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 16*(3), 58–73.
Tiwari, S., Jain, A., Ahmed, N. M. O. S., Alkwai, L. M., Dafhalla, A. K. Y., & Hamad, S. A. S. (2021). Machine learning-based model for prediction of power consumption in smart grid-smart way towards smart city. *Expert Systems, e12832*.
Tiwari, S., Jain, A., Sharma, A. K., & Almustafa, K. M. (2021). Phonocardiogram signal based multi-class cardiac diagnostic decision support system. *IEEE Access, 9*, 110710–110722.
Vikas, B., Anuhya, B. S., Chilla, M., & Sarangi, S. (2018). A Critical Study of Polycystic Ovarian Syndrome (PCOS) Classification Techniques. *International Journal of Computational Engineering & Management, 21*(4), 1–7.
Wang, D., Li, R., Wang, J., Jiang, Q., Gao, C., Yang, J., … Hu, Q. (2020). Correlation analysis between disease severity and clinical and biochemical characteristics of 143 cases of COVID-19 in Wuhan, China: A descriptive study. *BMC infectious diseases, 20*(1), 1–9.
Wisesty, U. N., & Nasri, J. (2016 August). In *Modified backpropagation algorithm for polycystic ovary syndrome detection based on ultrasound images* (pp. 141–151). Cham: Springer.
Wong, G. M., Lewis, J. M., Knudson, C. A., Millan, M., McAdam, A. C., Eigenbrode, J. L., ... & House, C. H. (2020). Detection of reduced sulfur on Vera Rubin ridge by quadratic discriminant analysis of volatiles observed during evolved gas analysis. *Journal of Geophysical Research: Planets, 125*(8), e2019JE006304.
Xie, N. N., Wang, F. F., Zhou, J., Liu, C., & Qu, F. (2020). Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network. *BioMed Research International, 2020*, 1–13.
Yu, D., Liu, Z., Su, C., Han, Y., Duan, X., Zhang, R., … Xu, S. (2020). Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier. *Thoracic Cancer, 11*(1), 95–102.
Zhang, X., Liang, B., Zhang, J., Hao, X., Xu, X., Chang, H. M., … Tan, J. (2021). Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening. *Molecular and Cellular Endocrinology, 523*, 111139.
Zhou, D., Xi, B., Zhao, M., Wang, L., & Veeranki, S. P. (2018). Uncontrolled hypertension increases risk of all-cause and cardiovascular disease mortality in US adults: The NHANES III Linked Mortality Study. *Scientific Reports, 8*(1), 1–7.
Zigarelli, A., Jia, Z., & Lee, H. (2022). Machine-Aided Self-diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study. *JMIR Formative Research, 6*(3), e29967.