Health and Technology (2022) 12:137–150
https://doi.org/10.1007/s12553-021-00613-y

ORIGINAL PAPER

Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques

Homay Danaei Mehr[1] · Huseyin Polat[1]

Received: 27 July 2021 / Accepted: 10 October 2021 / Published online: 28 October 2021

Abstract
Polycystic ovary syndrome (PCOS) has been determined as one of the serious health problems among women which affects women's fertility and leads to crucial health conditions. Hence, early diagnosis of polycystic ovary syndrome can be effective in the treatment process. Recently, machine learning methods have acquired promising results in medical diagnosis. Furthermore, feature selection techniques which generate the most significant subset of features, can reduce the computational time and improve the performance of classifiers. Conventional single machine learning algorithms classify datasets in a single process with an individual model whereas ensemble machine learning algorithms create multiple process with a combination of two or more models which can achieve more accurate results. Therefore, considering the advantages of ensemble classifiers and feature selection methods, in this study, traditional and ensemble classifiers were applied on the Kaggle PCOS dataset to diagnose polycystic ovary syndrome. Furthermore, the performance of various classifiers (i.e., Ensemble Random Forest, Extra Tree, Adaptive Boosting (AdaBoost) and Multi-Layer Perceptron (MLP)) were investigated using the dataset with all features and reduced subsets of features which were generated by filter, embedded and wrapper feature selection methods. The experimental results demonstrated that the feature selection methods had beneficial effects on the improvement of the performance of all classifiers. Moreover, Ensemble Random Forest classifier by using the reduced subset of features based on the embedded feature selection method surpassed other classifiers with Accuracy of 98.89% and Sensitivity of 100% in this study and other studies in the literature.

Keywords Polycystic ovary syndrome diagnosis · Feature selection · Machine learning · Random forest · Ensemble learning

1 Introduction

Polycystic ovary syndrome (PCOS) is one of the most prevalent endocrine disorders which affects 8–13% of women in productive age and 6–18% of younger females [1]. Moreover, approximately 70% of PCOS diagnosed women are infertile due to the accumulation of various cysts in their ovary which leads to ovulation failure. Although, geographic location, and genes are the major causes of this disorder, unhealthy diet, infectious diseases can aggravate the condition [2].

Not only do PCOS-patients suffer from infertility, but they also experience symptoms of imbalanced female hormones, high levels of male hormones and hair loss. Furthermore, PCOS can result in other serious disorders such as high blood pressure, mental disorders, heart diseases, diabetes (type 2) and endocrine disorders, as well. Hence, early prognosis of PCOS which mostly relies on both physical symptoms (e.g., abnormal hair growing under females' chin) and biochemical and clinical examinations is an essential issue and it can be beneficial in the treatment process [3, 4].

Since the advent of various technologies in biomedical and healthcare has resulted in acquisition of huge amount of data, extraction of rational conclusions to identify diseases has become a challenging task. Therefore, machine learning techniques as the sub-branch of artificial intelligence methods can learn relations and patterns among data to make logical decisions about unseen data in different fields of study particularly diagnosis of diseases such as diabetes, infectious diseases, Autism, etc. [5–7], in order to assist physicians and accelerate the process of diagnosis and prediction.

✉ Homay Danaei Mehr
h.danai.mehr@gmail.com

[1] Graduate School of Natural and Applied Sciences, Gazi University, Teknikokullar, 06500 Ankara, Turkey

Furthermore, as recently machine learning techniques have achieved satisfying results in terms of diagnosis of diseases which contributes to early treatment and reduction of death toll, scientists have been inclined to take the advantage of different machine learning methods to predict diseases [8].

Classification methods which can discriminate different categories, are the most prevalent machine learning techniques in medical diagnosis. However, the accuracy of classifiers can be affected by high dimensional data, due to the overfitting conditions and costly computational tasks. Thus, selecting the most significant data can reduce the overfitting risk and improve the processing time and accuracy of the classification methods [9].

Although different studies have focused on detection of follicle and classification of PCOS in women using ultrasound images [10, 11], in general the first stage of prognosis of PCOS is screening patients based on clinical examinations. Hence, in this study, to diagnose PCOS, different machine learning algorithms (i.e., Multi-Layer Perceptron (MLP), Ensemble Random Forest (RF), Ensemble Boosting and Ensemble Extra Tree) were applied on Kaggle PCOS clinical dataset [12]. Moreover, the accuracy rate of each classifier was evaluated after applying various feature selection methods (i.e., wrapper, filter and embedded methods).

The rest of the paper is organized as follows: a literature review of other studies which have used various machine learning algorithms and feature selection methods to diagnose PCOS using the Kaggle PCOS dataset and other different datasets, is presented in the second section; a brief explanation of the used classification algorithms and feature selection methods as well as dataset description are presented in the third section. Additionally, the description of classification and feature selection processes which were applied in this study were given in the fourth section. The experimental results and discussion of the used methods and conclusion were presented in the fifth and sixth sections, respectively.

2 Literature review

In this section a brief description of studies which applied machine learning methods and feature selection techniques to classify different clinical PCOS datasets, are presented.

Mehrotra et al. [13] evaluated the performance of LR and Bayesian classifiers in terms of classification of selected features of PCOS dataset which was collected from Ghosh Dastidar Institute for Fertility Research (GDIFR), Kolkata. Furthermore, statistical methods of t-value and p-value were used as feature selection methods. Experimental results demonstrated that Bayesian classifiers achieved higher accuracy rate (93.93%) than LR [13]. Meena et al. [15] proposed Neural Fuzzy Rough Subset Evaluating (NFRSE) as a feature selection method and used Information Gain Subset Evolution (IGSE) to provide selected features of their PCOS patients' dataset [14] for ID3 and J48 decision tree classifiers. Results showed that NFRSE-ID3 obtained less error rate than the others [15]. Meena et al. [15] evaluated the performance of their proposed hybrid feature selection and classification method (a combination of Neural Fuzzy Rough Set (NFRS) and ANNs) and various feature selection methods (PCA, Gain Ration, Information Gain and Correlation based Feature Selection (CFS)) and classifiers (SVM, ANNs, Decision Tree and NB) in classification of their PCOS dataset. Experimental results demonstrated that the hybrid method could achieve the highest classification accuracy (83.83%) among other methods [16]. Balogun et al. [17] used C4.5 Decision Tree, NB and MLP to classify PCOS dataset of Obafemi Awolowo University and the results presented that C4.5 Decision Tree and MLP surpassed NB, by the accuracy rate of 74.359 [17]. Vikas et al. [19] compared performance of three classifiers (NB, Decision Tree and ANNs) in classification of the PCOS survey dataset [18] and the results showed that NB obtained higher accuracy rate (97.65%) than the other two classifiers [19]. Denny et al. [20] applied Principal Component Analysis (PCA) to identify optimal features of their PCOS dataset which was gathered from infertility treatment centers at Thrissur and applied various classification methods (i.e., CART, KNN, Gaussian NB, RF, SVM and LR) to early diagnosis of PCOS. Results demonstrated the Superiority of RF (by 89.02% accuracy rate) over other classifiers [20]. Bharati et al. [21] applied filter based univariate feature selection method to select the most important features using Kaggle PCOS dataset. Furthermore, they used gradient boosting, RF, Logistic Regression (LR) and Hybrid Random Forest and Logistic Regression (RFLR) to classify PCOS using selected ten features. The results demonstrated that RFLR could surpass other classifiers in terms of the accuracy of classification (91.01%) and recall value (90%) [21]. Hassan and Mirza [22], used different machine learning methods (i.e., RF, Support Vector Machine (SVM), CART, Naïve Bayes (NB) and LR) to classify Kaggle PCOS dataset. Moreover, experimental results presented that RF could outperform other classifiers by achieving 96% classification accuracy rate [22]. Neto et al. [23], compared the performance of different classifiers (i.e., SVM, MLP, RF, LR and Gaussian NB) in classification of PCOS Kaggle dataset. RF could surpass other classifiers by achieving the accuracy rate of 95% and a precision of 96% [23]. Munjal et al. [24] used three machine learning methods (i.e., Extra Trees, RF and Decision Tree) and Genetic Algorithms (GA) as a feature selection method to diagnose PCOS using Kaggle PCOS dataset with nine features. Results showed that Extra Trees could surpassed the other two classifiers with the highest accuracy rate (88%) [24]. Nandipati et al. [26] compared different classification methods and feature selection methods in Python and Rapid Minder tools to classify Kaggle PCOS dataset. Their results showed that RF (using RapidMiner) without feature selection could achieve accuracy rate of 93.12% and KNN and SVM gained 90.83 accuracy rate using ten selected features [25]. Prapty and Shitu [25] applied Decision Tree to select effective features of their PCOS dataset and they also applied RF, SVM, NB and KNN to classify the dataset. Finally, results indicated that RF could outperform other classifiers by 93.5% accuracy rate [26]. Pushkarini and Anusuya [27] applied RF, Linear Regression and K-Nearest Network (KNN) methods on a collected PCOS dataset from infertility clinic and research center in order to classify the dataset. Experimental results demonstrated that RF could achieve the highest R^2 compared to the other two methods [27]. Tanwani [28] applied filter methods to find correlation between features of Kaggle PCOS dataset and classified the dataset using LR and KNN. Results showed that LR achieved higher accuracy rate by 92% (with ten selected features) compared to KNN [28]. Thomas and Kavitha [29] applied hybrid classification method (combination of ANNs and NB) to classify clinical PCOS dataset which was gathered from hospitals and scanning center at Thodupuzha. Moreover, their hybrid classifier could obtain 95% accuracy rate [29]. Inan et al. [30] used statistical methods (i.e., Chi-Square and Analysis of Variance (ANOVA)) for feature selection of Kaggle PCOS dataset to provide inputs for different classifiers (i.e., SVM, KNN, RF, Ensemble Adaptive Boosting (AdaBoost), XGBoost, NB and MLP). Experimental results showed that XGBoost could achieve higher classification accuracy rate (95.83) compared to the other classifiers [30]. Zhang et al. [31] applied XGB, KNN, RF and stacking classifier model including KNN, RF and XGB in its first layer and XGB in its second layer for classification of two datasets obtained from Raman spectra of the Shengjing Hospital of China Medical University. Additionally, the dataset included two groups of data from follicular fluid samples and the plasma samples. Experimental results indicated that the stacking classifier model achieved higher accuracy rate (89.32%) by using follicular fluid than using plasma samples and it could surpass other classifiers as well [31].

3 Materials and methods

Different classification algorithms (i.e., Ensemble RF, Ensemble AdaBoost, ANNs (MLP) and Ensemble Extra Tree) and also filter, wrapper and ensemble feature selection methods which were used to diagnose PCOS, are explained in this section.

3.1 Artificial neural networks (ANNs)

Artificial Neural Networks (ANNs) [32] which are inspired by the networks of the nerve cells of human brain, include artificial interconnected neurons imitating the process of biological central nervous system. The structure of ANNs is based on the modification and adjustment of weights during training of the network. Additionally, after calculation of the output based on the class labels, weights will be re-calculated by considering the differences between the actual class labels and the predicted output labels. One of the prevalent types of ANNs is Feed Forward Neural Networks in which the calculated weight values of each previous layer is sent to the next layer.

Multi-Layer Perceptron (MLP) is one of the most significant classes of Feed Forward Neural Network which uses Back Propagation algorithm in training process to reduce the error rate of weight adjustment of the network [33] and has achieved satisfying results in classification of different biomedical datasets [34, 35].

3.2 Ensemble learning approach

In Ensemble learning the combination of classifiers are aggregated to generate a single improved predictive model which makes decision based on the classification results of all individual classifiers. As a result, it usually can achieve higher accuracy rate than individual classifiers [36]. Ensemble AdaBoost, RF and Extra Trees which were used in this study as three types of ensemble learning models, are described below.

3.2.1 Random forest (RF)

RF which was introduced by Breiman [37] is based on the combination of various decision trees. This Ensemble supervised method can decrease the prediction error rate of the classification task due to the use of bootstrap aggregation and bagging to select a sample. Therefore, after arbitrary selection of x numbers of features among all features, a node is chosen among x features as a candidate split node. Subsequently, feature selection is done at each division and the splitting process will stop when decision tree is completed. The capability of RF in dealing with high dimensional data and various types of data particularly nonparametric data makes it as a desirable and efficient method [38].

3.2.2 Ensemble adaptive boosting (AdaBoost)

AdaBoost [39] which is one of the most common ensemble boosting algorithms, takes the advantage of using the combination of independent singular hypothesis to boost the accuracy by improving the performance of each individual weak learners. During training process, it changes the distribution of training set among individual classifiers to minimize the training error and after the completion of training process in each iteration, weights are adjusted. Consequently, for each correctly classified instance, weights will be reduced and final decision will be made by considering the result of all single classifiers. Furthermore, boosting algorithms are less vulnerable to overfitting problem.

Various machine learning classification algorithms such as Decision Tree, SVM or any other algorithms can be used as a base classifier of AdaBoost [40].

3.2.3 Ensemble extra tree

The Extra Tree which is also known as Extremely Randomized Tree, is a type of ensemble machine learning algorithm generating numerous unpruned decision trees among training set alongside top-down procedure. Contrary to the RF, it randomly selects features and cut-points through the process of splinting the nodes. Moreover, not only does Extra Tree differ from other ensemble-based trees in terms of randomly selecting the cut-points, but it also uses the entire training set to expand the tree which can result in efficient reduction of variance. Therefore, the final decision of the classifier is made based on majority voting [41].

3.3 Feature selection

Feature selection method is applied to reduce features' dimensions as well as to select a subset of dataset in which the distribution of classes does not differ from the original datasets. Consequently, through feature selection, not only the most significant and relevant features to the main class/classes are defined, but also irrelevant features are removed. However, attempting to create the smallest subset can deceive the classifier and result in unrealistic results. As a result, selecting the most effective subset is a crucial issue. Furthermore, feature selection is beneficial in terms of size reduction of data and storage space, shorter training time, improvement of the accuracy, and reduction of overfitting problem. Thus, the major purpose of feature selection is creating the smaller subset from the original dataset which determine the main and pivotal features to facilitate classification process and improve the performance of the classifiers as well as interpretation of features. Feature selection methods are categorized in three groups, filter, wrapper and embedded methods (Fig. 1) which generate subset of features from original dataset and evaluate the merit of the subset by considering their own criteria to provide the most optimal feature set [42, 43].

However, the combination of feature selectin methods can be developed as hybrid models to take the advantage of the efficiency of each method in creating an appropriate feature set. Each feature selection methods are discussed below.

3.3.1 Filter method

Filter method which is known as the earliest feature selection methods, selects features by considering their characteristics such as distance or their rank among others, rather than using machine learning algorithms. Although filter methods use statistical techniques to select and evaluate the subset which make them scalable and swift, they ignore dependency among features and the interaction with classifiers (Fig. 2) [44]. In this study Pearson method was selected as a filter feature selector, since it could improve the accuracy of

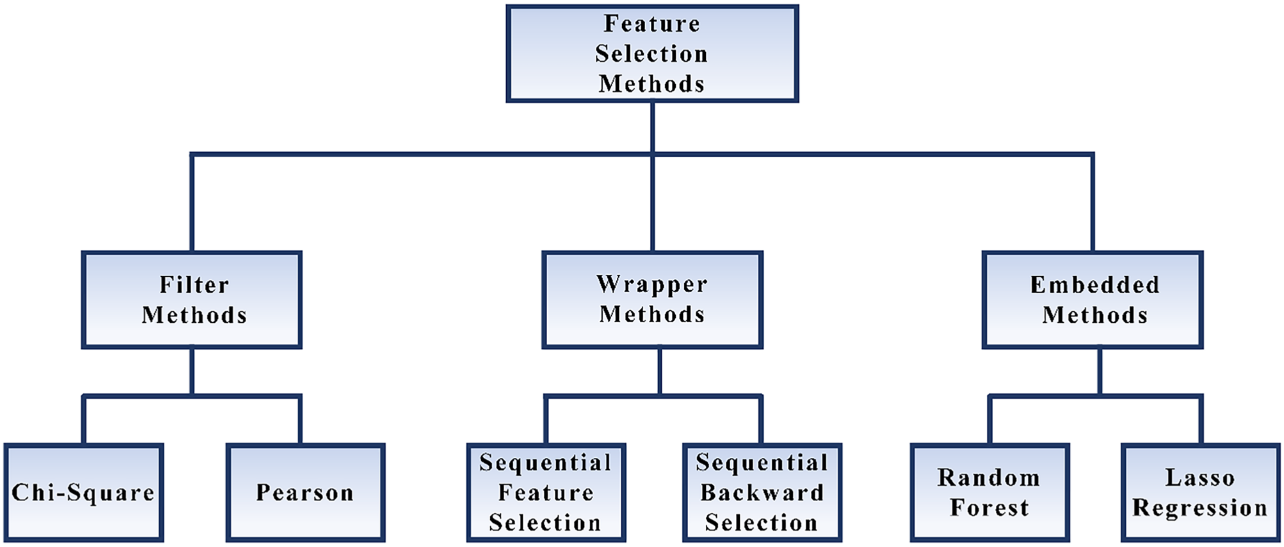

Fig. 1 Different kinds of feature selection methods with some examples

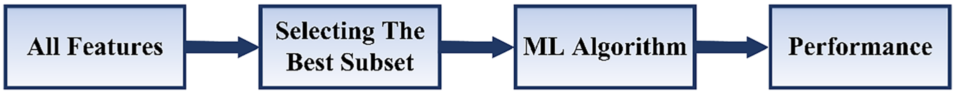

Fig. 2 Filter feature selection method

classifiers on PCOS dataset and it was faster than other filter methods such as Information Gain and Chi Square, etc. The process of Pearson method is described below.

The Pearson feature selection method is a kind of statistical filtering methods which measures the correlation of features ranging between –1 and 1. 1 indicates high correlation and –1 indicates low correlation of two features. Since Pearson method is efficient in computing correlation of features compared to other methods, it is considered as the most popular filter method in various studies [45, 46].

3.3.2 Wrapper method

Wrapper Method is known as close-loop feature selection method. In this method, feature selection algorithm is wrapped around the machine learning algorithm and the efficiency of the feature selection algorithm is assessed by accuracy or error rate of the classifier. Thus, the most optimal subset will be selected when the error rate of the classifier during feature selection process is not conspicuous indicating that the wrapper method depends on the learning algorithm to select the most efficient subset of features. The paramount advantage of the wrapper method over filter method is achieving higher accuracy due to its dependency on the learning algorithm. However, it includes complex computation and also it can confront overfitting problem (Fig. 3) [47]. In this study sequential backward selection algorithm (SBS) was selected as a wrapper method due to its ability which led to be adjusted to the PCOS dataset and decreased the validation error and could avoid excessive overfitting in each iteration compared to the other wrapper methods. The process of SBS is described below.

In the process of collecting the effective features, the importance score of each feature is determined by considering the full dataset at the first step and in each iteration features with the least importance score are eliminated. Hence, the performance of the model will be improved at each iteration. After merging remained features, the round will commence until no more improvement is recognized by eliminating features [48].

3.3.3 Embedded method

In embedded method which is known as built-in- feature selection method, feature selection is embedded as a component of the learning algorithm to assist the feature evaluator in the process of selecting the optimal subset. The most common learning algorithms in the embedded methods are various types of decision tree and ANNs. Moreover, as embedded method does not repeatedly apply classifier to evaluate each feature of the selected subset, it does not include complicated computation. As a result, its less computational process relies on the selection of features through implementing learning algorithm which makes it more effective feature selection mechanism compared to filter and wrapper methods (Fig. 4) [49]. In contrast to some embedded feature selectors such as Lasso and Ridge Regression using penalty function to reduce overfitting problem which increases the computational cost, RF has the capability to deal with overfitting problem without using penalty function which makes it conspicuously efficient and convenient in solving various problems [50, 51]. Hence, in this study RF feature selection is selected as an embedded feature selector due to its advantage in dealing with overfitting problem on PCOS dataset compared to other embedded feature selection methods. The process of RF feature selection method is explained below.

Firstly, bagging method is applied on the training set to generate different subsets of features which are used to construct various decision trees. Moreover, in the growing process of

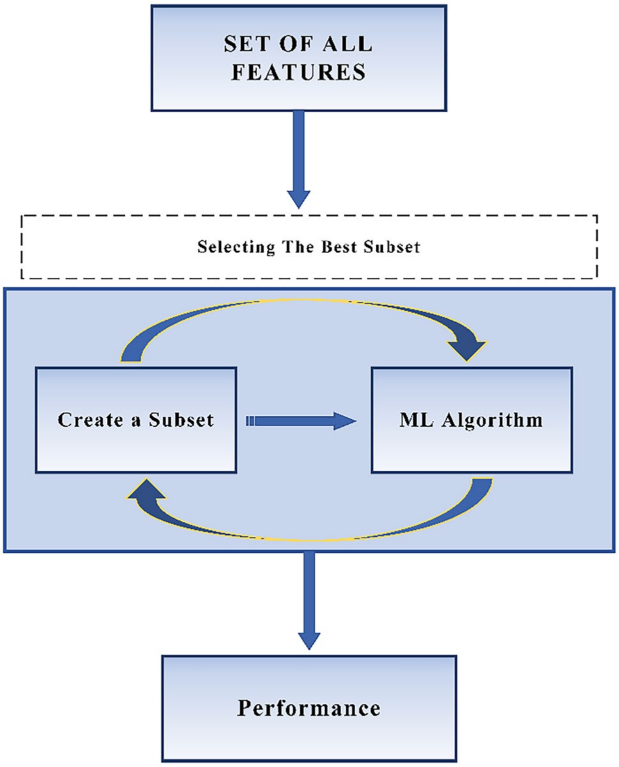

Fig. 3 Wrapper feature selection method

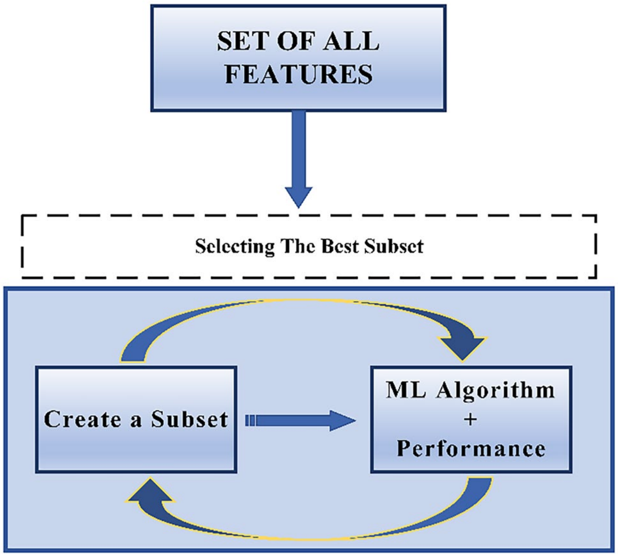

Fig. 4 Embedded feature selection

each tree, splitting of each node relates to the selected features among candidate subsets which are arbitrarily selected. Therefore, splitting of nodes results in the growth of all trees without applying pruning and each tree is considered as a principal classifier. In the final step, all trees are connected to produce a Random Forest which generates the efficient subset of features [50].

3.4 Performance metrics

In this section the most common evaluation metrics [52] which are considered in this study to investigate the performance of classifiers are described below.

True Positive (TP): Samples with positive label are correctly predicted as positive.
True Negative (TN): Samples with negative label are correctly predicted as negative.
False Positive (FP): Samples with negative labels are inaccurately classified as positive.
False Negative (FN): Samples with positive labels are inaccurately classified as negative.

Accuracy: Indicates the total proportion of correctly predicted classes.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Sensitivity (Recall): Indicates what proportion of predicted class labels as positive labels (with-PCOS labels) belong to positive labels.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

Specificity: demonstrates what proportion of predicted class labels as negative labels (without-PCOS labels) belong to negative labels.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

Precision: demonstrates the proportion of predicted classes as "with-PCOS" labels on all positive classified labels.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{F} - \text{measure} = 2 \times \frac{\text{Sensitivity} \times \text{Precision}}{\text{Sensitivity} + \text{Precision}} \tag{5}$$

Receiver Operating Characteristic (ROC) Curve: Is one of the evaluation tools of performance analysis of the classifier in which the X-axis represents the FP-rate and Y-axis represents TP-rate. The area under the ROC curve (AUC) which is between 0 (all classes are predicted inaccurately) and 1 (all class labels are accurately classified), interprets the performance of the classifier.

3.5 Dataset and preprocessing

In this study PCOS dataset which was gathered from 10 different Indian hospitals, is collected from Kaggle Dataset Repository [12]. The dataset includes 43 features which are based on physical and medical examination from 541 women and the class feature indicates which person is diagnosed with PCOS (177 instances) or without PCOS (364 instances).

Since dataset was provided in two sections, in preparation process of the dataset, two separated datasets were merged based on patient number. Moreover, some features which include missing values (i.e., Marriage Status, beta-HCG, AMH and Fast food) were filled with median value of the instances. Consequently, after elimination of patient number, the prepared dataset contains 42 features without missing values. Additionally, feature descriptions are depicted in Table 1.

4 Classification of PCOS

In order to diagnose patients with PCOS or without PCOS using machine learning techniques, the prepared dataset was divided into train and test sections using tenfold cross validation.

Table 1 Feature Description of PCOS dataset

No	Feature	No	Feature	No	Feature
1	PCOS class label	15	I beta-HCG	29	Weight gain
2	Age	16	II beta-HCG	30	Hair growth
3	Weight	17	FSH: Follicle stimulating hormone	31	Skin darkening
4	Height	18	LH: Luteinizing Hormone	32	Hair loss
5	BMI: Body Mass Index	19	FSH/LH	33	Pimples
6	Blood Group	20	Hip (inch)	34	Fast food
7	Pulse rate	21	Waist (inch)	35	Regular Exercise
8	RR (breaths/min)	22	Waist: Hip Ratio	36	BP Systolic
9	Hb: hemoglobin	23	TSH: Thyroid-Stimulating Hormone	37	BP Diastolic
10	Cycle: menstrual cycle	24	AMH: Anti-Müllerian Hormone	38	Follicle No. (L)
11	Cycle length (days)	25	PRL: Prolactin	39	Follicle No. (R)
12	Marriage Status	26	Vitamin D3	40	Avg. F size (L)
13	Pregnant	27	PRG: Progesterone	41	Avg. F size (R)
14	No. of abortions	28	RBS: Random Blood Sugar	42	Endometrium

Train dataset went through four classification methods (i.e., MLP (as Feed Forward ANNs), Ensemble AdaBoost, Ensemble RF and Ensemble Extra Tree) in two steps including with and without feature selection methods.

Since each patient can have different symptoms and/or combination of symptoms, diagnosis process may mostly depend on the most significant features. Therefore, in order to select the most effective features and reduce dimension of the dataset to assist physicians during diagnosis process, feature selection methods can play a pivotal role.

In this study, to select important features, various feature selectin methods were applied on the dataset. Thus, Sequential Backward Selection (SBS) as a wrapper feature selection method, Pearson method as a filter feature selection method and RF as an embedded feature selection method, were applied on the dataset to reduce the dimension of the dataset and to select the most significant features. The framework of this study including all steps in diagnosis of PCOS is illustrated in Fig. 5.

The performance of classification algorithms on both datasets including dataset with whole features and datasets with reduced features was evaluated. Hence, in the first step, dataset was classified by five classification algorithms and the performance of each classifier was calculated. Furthermore, Ensemble RF, MLP, Ensemble AdaBoost and Ensemble Extra Tree classifiers were applied on each reduced dataset resulted by different feature selection methods. In this study Python as a programming language, Scikit-learn library and Jupiter notebook were used to develop machine learning and feature selection models.

5 Experimental results and discussion

In this study three types of feature selection methods including Sequential Backward Selection (SBS) as a wrapper feature selection method, Pearson method as a filter feature selection method and RF as an embedded feature selection method were applied on the PCOS dataset to evaluate the effect of feature selection on the performance of various classifiers in recognition of patients with and without PCOS. Feature selection methods reduced the number of features of the original dataset which included 42 features. Pearson filter feature selection method decreased the number of features to 33 and SBS wrapper feature section method reduced the number of features to 30. Furthermore, the number of features of the PCOS was reduced to 28 by applying RF embedded feature selection.

Moreover, different classifiers were applied on the dataset with whole features (without applying feature selection) and datasets with reduced features resulted by feature selection methods.

Confusion matrix of various classifiers using different feature selection methods on the PCOS dataset with whole features and reduced features are depicted in Tables 2, 3, 4, and 5.

Although correctly classified labels with and without PCOS in Ensemble RF classifier (with TN = 347 out of 364 and TP = 173 out of 177) and MLP (with TN = 346 and TP = 174) are not exactly the same, the differences in their FN and FP compensate the differences of TP and TN and they could show approximately the same performance (accuracy = 96.11%) in in classification of PCOS without using feature selection algorithms (Tables 2 and 6).

As a result, among four classifiers in classification of PCOS with whole features, Ensemble RF and MLP could classify higher number of patients with PCOS and without PCOS and their inaccurately classified instances (FP and FN) are less than Ensemble AdaBoost and Ensemble Extra Tree. Moreover, Ensemble Extra Tree classifier with the highest inaccurately classified instances (FN = 9 and FP = 25) and the lowest accurately classified instances (TN = 339 and

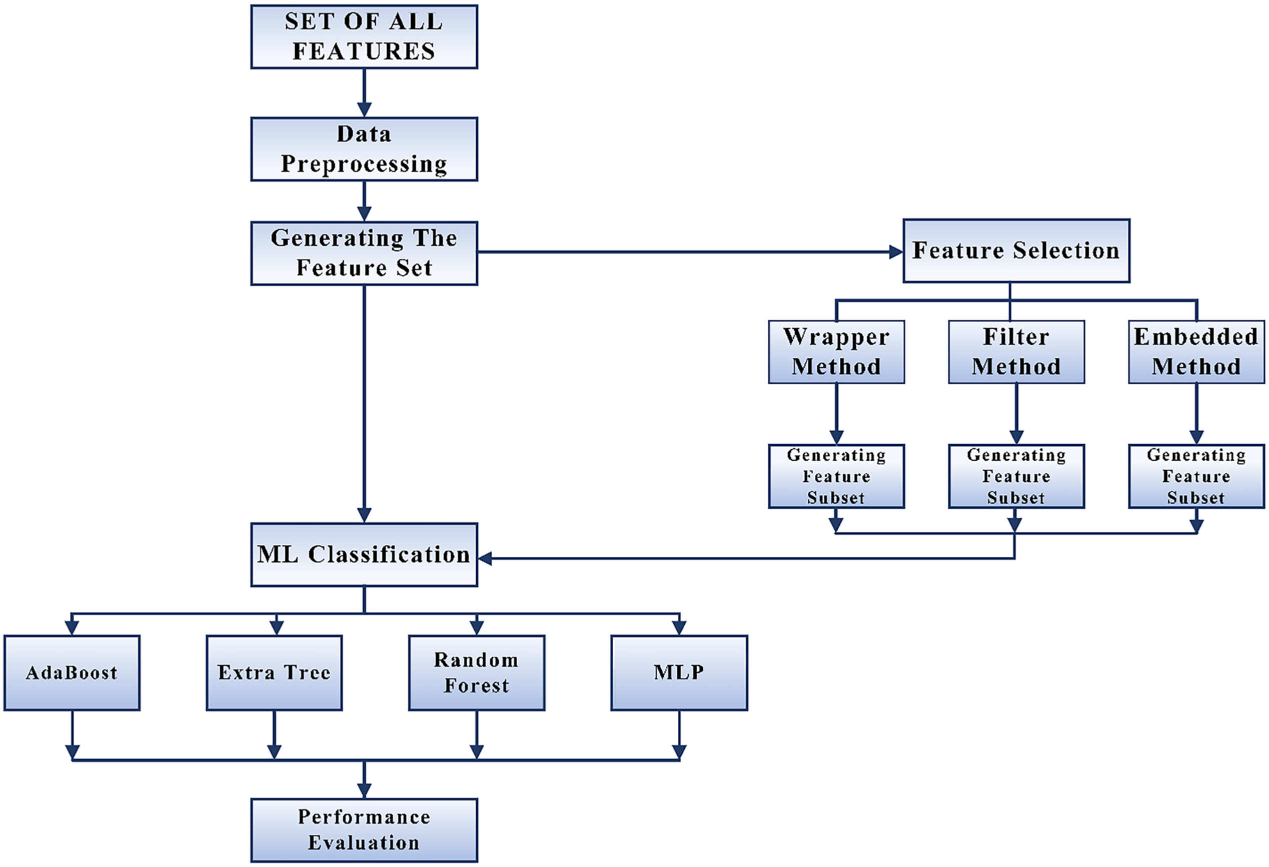

Fig. 5 Framework of the study

TP = 168) showed the lowest performance among all classifiers in classification of the dataset without using feature selection methods (Table 2). Considering the number of accurately and inaccurately predicted classes on PCOS dataset without using any feature selection method, MLP and RF could acquire the highest and the same classification accuracy rate (96.11%). Furthermore, the less value of FP and FN the classifiers acquire, the more value of sensitivity, specificity and precision are achieved. Consequently, the value of specificity (95.32%) and precision (91.05%) of the Ensemble RF classifier is higher than the specificity (95.05%) and precision (90.15%) of MLP due to its low value of FP. Additionally, the highest value of sensitivity (98.30%) of MLP is due to its lowest value of FN (Table 6).

Comparison of the results of correctly and incorrectly classified instances demonstrates that all feature selection

Table 2 Confusion Matrix of all classifiers on whole dataset without using a feature selection method

Classifier	Actual	Prediction	
		Without PCOS	With PCOS
Ensemble Random Forest	Without PCOS	347	17
	With PCOS	4	173
MLP	Without PCOS	346	18
	With PCOS	3	174
Ensemble Extra Tree	Without PCOS	339	25
	With PCOS	9	168
Ensemble AdaBoost	Without PCOS	347	17
	With PCOS	6	171

Table 3 Confusion Matrix of all classifiers on reduced dataset using Pearson filter feature selection

Classifier	Actual	Prediction	
		Without PCOS	With PCOS
Ensemble Random Forest	Without PCOS	348	16
	With PCOS	2	175
MLP	Without PCOS	348	16
	With PCOS	3	174
Ensemble Extra Tree	Without PCOS	340	24
	With PCOS	7	170
Ensemble AdaBoost	Without PCOS	346	18
	With PCOS	4	173

Table 4 Confusion Matrix of all classifiers on reduced dataset using SBS wrapper feature selection

Classifier	Actual	Prediction	
		Without PCOS	With PCOS
Ensemble Random Forest	Without PCOS	354	10
	With PCOS	1	176
MLP	Without PCOS	350	14
	With PCOS	2	175
Ensemble Extra Tree	Without PCOS	343	21
	With PCOS	5	172
Ensemble AdaBoost	Without PCOS	348	16
	With PCOS	3	174

methods could have positive impact on the performance of the classifiers.

Thus, Pearson filter method which reduced the number of features of the dataset to 33, could improve the classifiers' performance in terms of increasing the number of correctly classified instances (TN and TP) and reducing the number of incorrectly classified instances (FN and FP). Therefore, among all classifiers which were applied on the reduced number of features by Pearson method, RF could take the advantage of the Pearson method by increasing the number of accurately predicted classes (TN = 348 and TP = 175) and decreasing the number of inaccurately predicted classes (FN = 2 and FP = 16). Moreover, the improvement of the result of classified instances in Ensemble Extra Tree using filter method is the lowest among other classifiers (Table 3).

Although Pearson filter method couldn't alter the overall accuracy rate (95.74%) of Ensemble AdaBoost compared to its performance in using the whole dataset, it could improve the accuracy rate of other classifiers. Therefore, Ensemble RF classifier could obtain the highest accuracy rate (96.67%) on the reduced number of features using Pearson method among all other classifiers (Table 7) due to its the highest value of TP and TN in prediction of classes with-PCOS and without-PCOS labels. Additionally, due to the lowest value of FN of the Ensemble RF classifier, its sensitivity (98.87) is higher than its competitors (Tables 3 and 7).

Table 5 Confusion Matrix of all classifiers on the reduced dataset using RF embedded feature selection

Classifier	Actual	Prediction	
		Without PCOS	With PCOS
Ensemble Random Forest	Without PCOS	358	6
	With PCOS	0	177
MLP	Without PCOS	356	8
	With PCOS	0	177
Ensemble Extra Tree	Without PCOS	344	20
	With PCOS	4	173
Ensemble AdaBoost	Without PCOS	352	12
	With PCOS	3	174

Table 6 Performance of classifiers without feature selection

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-measure (%)
Ensemble Random Forest	96.11	97.74	95.32	91.05	94.27
MLP	96.11	98.30	95.05	90.15	94.04
Ensemble AdaBoost	95.74	96.61	95.32	90.95	93.69
Ensemble Extra Tree	93.71	94.91	93.13	87.04	90.80

Applying SBS wrapper feature selection method which reduced the number of features to 30, could assist classifiers to boost the number of correctly predicted classes and reduce the number of incorrectly predicted classes. Consequently, the most conspicuous improvement in the number of accurately classified classes (TN and TP) and reduction in the number of inaccurately predicted classes (FN and FP) is obtained by Ensemble RF classifier (FN = 1 and FP = 10) and MLP (FN = 2 and FP = 14) (Table 4).

Since SBS wrapper method could assist classifiers to make progress in prediction of class labels and decrease their deficiency, all classifiers could achieve higher accuracy rate and compared to their performance on the dataset with all features. Hence, the accuracy rate of RF classifier is the highest (97.96%) and the accuracy rate of Ensemble Extra Tree is the lowest (95.19%) among other classifiers using the reduced features resulted by SBS wrapper method (Table 8). Furthermore, due to the highest values of TN and TP and the lowest values of FP and FN in Ensemble RF classifier, it could achieve the highest values of sensitivity, specificity, precision and F-measure. (Tables 4 and 8).

Table 7 Performance of classifiers using reduced features resulted by Pearson filter method

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-measure (%)
Ensemble Random Forest	96.67	98.87	95.60	91.62	95.10
MLP	96.48	93.30	95.60	91.57	94.81
Ensemble AdaBoost	95.74	97.74	95.05	90.57	94.01
Ensemble Extra Tree	94.26	96.04	93.40	87.62	91.63

Table 8 Performance of classifiers using reduced features resulted by SBS wrapper method

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-measure (%)
Ensemble Random Forest	97.96	99.43	97.25	94.62	96.96
MLP	97.04	98.87	96.15	92.59	95.62
Ensemble AdaBoost	96.48	98.30	95.60	91.57	94.81
Ensemble Extra Tree	95.19	97.17	94.23	89.11	92.96

Furthermore, RF embedded feature selection method, which reduced the number of features to 28, could assist all classifiers to acquire the highest number of correctly predicted classes and the lowest number of incorrectly classified instances, compared to their own achievements in using other feature selection methods. Hence, Ensemble RF and MLP are the most prosperous classifiers which could correctly classify all patients with PCOS (TP = 177 and FN = 0) and reduced the number of incorrectly predicted classes without PCOS (FP) to 6 and 8 which indicates only 6 instances by Ensemble RF classifier and 8 instances by MLP classifier were not correctly classified as without PCOS labels. Additionally, AdaBoost using reduced features by embedded RF feature selection could correctly predict 352 (TN) classes as without-PCOS label and consequently it could reduce incorrectly classified classes as with-PCOS label (FP = 12) which are more satisfying than the results of Extra Tree (Table 5).

In addition, since RF embedded method could effectively improve the performance of all classifiers by considering their achievements of the highest values of TN and TP and the lowest values of FN and FP, the accuracy rate of all classifiers showed their highest values. Thus, Ensemble RF classifier using 28 features of the dataset could obtain higher accuracy rate (98.89%), specificity (98.35%), precision (96.72%) and F-measure (98.33%) compared to the other classifiers. Moreover, it obtained the same value of sensitivity (100%) with MLP due to the lowest value of FN (0) (Tables 5 and 9).

On the whole, Ensemble RF and MLP could acquire the highest and the same accuracy rate in diagnosis of PCOS without using any feature selection methods (Table 6). However, Ensemble RF could surpass all classifiers using reduced number of features resulted by different feature selection methods. Furthermore, the most promising accuracy rate of Ensemble RF classifier is 98.89% which indicates using reduced numbers of features achieved by RF embedded method could assist the classifier to decrease the rate of inaccurately classified instances to 1.11%. (Table 10).

Additionally, the ROC curve of the Ensemble RF classifier using reduced features resulted by RF embedded feature selection method which presented the highest performance and the ROC curves of the Ensemble Extra Tree without using any feature selector and using Pearson filter method which presented the lowest performance, are depicted in Fig. 6. Comparison Results demonstrate that since ROC curve of Ensemble RF classifier using RF embedded feature selection method is near to 1 and its AUC is higher than AUC of the weakest classifiers (i.e. Extra Tree using filter feature selection method and Extra Tree using all features without using a feature selector), it could perform perfectly at distinguishing classes with-PCOS and without-PCOS labels. In conclusion, between the both Extra Trees which resulted in the lowest performance, since AUC of ROC curve of Extra Tree using filter method is approximately larger than the AUC of Extra Tree using all features, its performance is better than Extra Tree using all features without using any feature selection method.

Not only did RF classifier achieve the highest performance in classification of PCOS dataset with both reduced number of features and all features of the dataset, but it also could achieve the highest accuracy rate among all studies in the literature in which the same dataset were used. Consequently, considering the previous studies in the literature using the Kaggle PCOS dataset, inordinate reduction of features couldn't distinctively improve the accuracy of classifiers and the highest accuracy (92.00%) among studies which reduced the number of features to 10 belongs to LR classifier [28]. Moreover, since the main purpose of feature selection is not only reducing the number of features, but it is also the improvement of the prediction ability of the classifier to gain more precise results and assist physicians. Consequently,

Table 9 Performance of classifiers using reduced features resulted by RF Embedded method

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-measure (%)
Ensemble Random Forest	98.89	100	98.35	96.72	98.33
MLP	98.52	100	97.80	95.67	97.78
Ensemble AdaBoost	97.22	98.30	96.70	93.54	95.86
Ensemble Extra Tree	95.56	97.74	94.50	89.63	93.50

Table 10 Summary of the performance of RF classifier on all features and reduced features of the dataset

Method	Number of features	Inaccurately classified instances	Accuracy rate
Without feature selection	42	3.89%	96.11%
Pearson filter feature selection	33	3.33%	96.67%
SBS wrapper feature selection	30	2.04%	97.96%
RF embedded feature selection	28	1.11%	98.89%

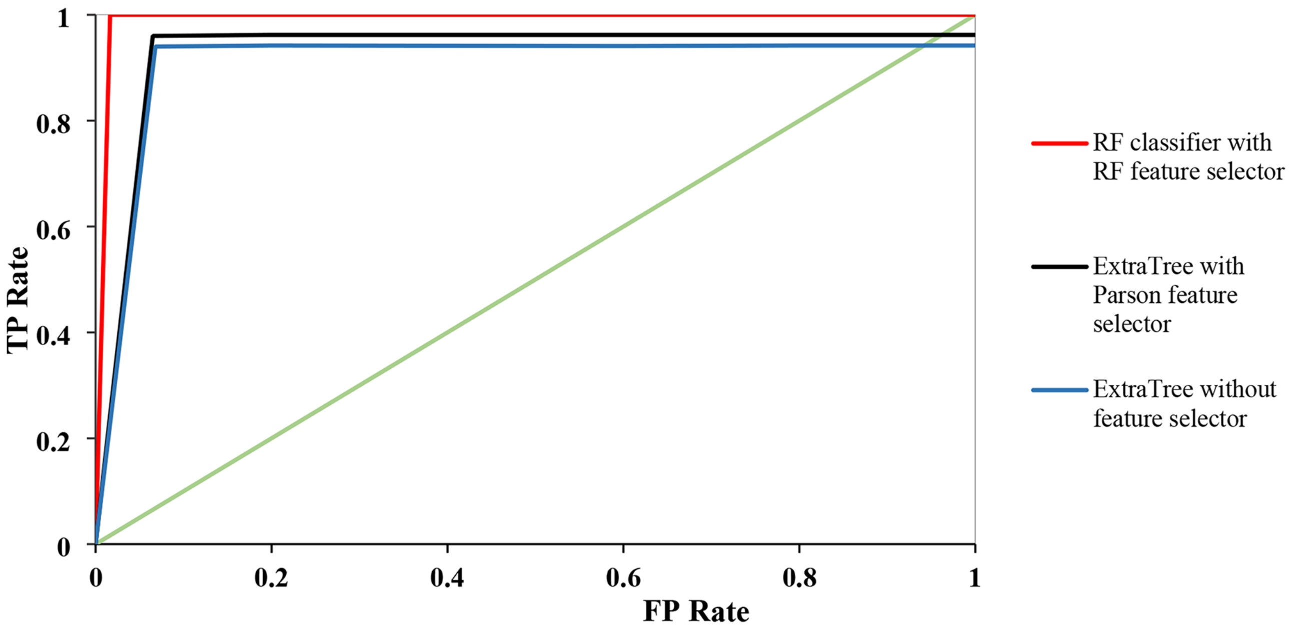

Fig. 6 ROC curve analysis of the classifiers with the highest and the lowest performance

Table 11 Comparative results of different methods in the literature on the same Kaggle PCOS dataset

ML algorithm	FS method	Number of selected features	Accuracy (%)	Reference
RF	GA	9	88	[24]
RF	-	-	93.12	[26]
KNN and SVM	-	10	90.83	
Hybrid RFLR	Filter based univariate	10	91.01	[21]
LR	filter based (correlation)	10	92	[28]
RF	-	-	95	[23]
XGBoost	Chi-Square and ANOVA	23	95.83	[30]
RF	-	-	96.11	This study
MLP	-	-	96.11	
RF	Pearson filter	33	96.67	
RF	SBS wrapper	30	97.66	
RF	RF embedded	28	98.89	

reducing 42 features to small numbers can cause critical situation in accurate diagnosis of PCOS. In [30] XGBoost classifier could achieve 95.83% accuracy rate on the 23 selected features using Chi-Square and ANOVA methods in classification of with-PCOS and without-PCOS classes which is the highest accuracy rate compared to other studies in the literature. However, in this study approximately most classifiers could acquire higher accuracy rate using different feature selection methods. Consequently, Ensemble RF classifier could obtain 98.89% accuracy rate on the 28 selected features resulted by RF embedded feature selection method which is the highest accuracy rate among all studies using the same dataset without taking the risk of ignoring significant features which can be crucial in diagnosis of PCOS (Table 11).

6 Conclusion

The main purpose of this study is to diagnose PCOS by applying different machine learning classification algorithms (i.e., Ensemble AdaBoost, Ensemble Extra Tree, Ensemble RF and MLP) on Kaggle PCOS dataset with whole features and also reduced features resulted by applying feature selection methods (i.e., Pearson filter method, RF embedded method and SBS wrapper feature selection method).

Although classifiers could obtain satisfying results in classification of classes with-PCOS and without-PCOS labels, feature selection methods, which attempt to select the most efficient subset of features, could assist classifiers to increase the number of correctly predicted instances (TN and TP) and reduce incorrectly predicted classes (FN and FP). Therefore, the performance of all classifiers using the reduced subset of features improved.

Pearson filter method reduced the number of features of the dataset to 33, SBS wrapper method selected 30 features among all features of the PCOS dataset and RF embedded feature selection method generated a subset with 28 features.

Among all classifiers Ensemble RF and MLP acquired the highest accuracy rate using the whole features. Ensemble RF classifier, however, could outperform other classifiers in terms of classification of PCOS dataset using the reduced subsets which were provided by all three feature selection methods. Moreover, the highest accuracy rate and sensitivity of Ensemble RF are 98.89% and 100% using the reduced subset of features generated by RF embedded feature selection method.

Not only should feature selection methods try to reduce the number of features to decrease the cost of classification algorithms, but also, they should provide a reliable subset with adequate amount of features to facilitate classification process to accurately predict class labels particularly in diagnosis process to assist physicians. Hence, in comparison to other studies in the literature which have used the same dataset, Ensemble RF classifier using different subsets resulted by various feature selection methods achieved the highest accuracy rate, particularly using 28 selected features resulted by RF embedded feature selection method.

Indeed, this method might not be an appropriate solution for classification of different datasets. However, for future study, proposing a hybrid feature selection method and classification algorithms are devised to evaluate their ability in classification of PCOS class labels alongside consulting an experienced physician to assess the validation of selected subset of features in terms of medical diagnosis.

Author contribution All the authors contributed equally in this research. Visualization, Investigation, Formal analysis and Writing of the original draft were performed by Homay Danaei Mehr. Conceptualization, Methodology, Editing and Review were conducted by Huseyin Polat.

Funding No funding was received for conducting this study.

Data availability The Polycystic Ovary Syndrome (PCOS) data which was used in this study is available at "https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos" and the processed dataset is available upon request.

Code availability Codes which were developed in Python are available upon request.

Declarations

Ethics approval Not Applicable. This paper does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate Not Applicable. This paper does not contain any studies with human participants or animals performed by any of the authors. Hence no informed consent is required.

Consent for publication Not Applicable. This paper does not contain any studies with human participants or animals performed by any of the authors. Therefore, no consent for publication is required.

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

1. Peña AS, Witchel SF, Hoeger KM, Oberfield SE, Vogiatzi MG, Misso M, Garad R, Dabadghao P, Teede H. Adolescent polycystic ovary syndrome according to the international evidence based guideline. BMC Med. 2020;18(72). https://doi.org/10.1186/s12916-020-01516-x.
2. Ajmal N, Khan SZ, Shaikh R. Polycystic Ovary Syndrome (PCOS) and genetic predisposition: A review article. Eur J Obstet Gynecol Reprod Biol. 2019;3: 100060. https://doi.org/10.1186/s12916-020-01516-x.
3. Soucie K, Samardzic T, Schramer K, Ly C, Katzman R. The diagnostic experiences of women with Polycystic Ovary Syndrome (PCOS) in Ontario, Canada. Qual Health Res. 2021;31(3):523–34. https://doi.org/10.1177/1049732320971235.

4. Zhu T, Cui J, Goodarzi MO. Polycystic ovary syndrome and risk of type 2 diabetes, coronary heart disease, and stroke. Diabetes. 2021;70(2):627–37. https://doi.org/10.2337/db20-0800.
5. Prasanth S, Thanka MR, Edwin ER, Ebenezer V. Prognostication of diabetes diagnosis based on different machine learning classification algorithms. Annals of R.S.C.B. 2021;25(5):372–95. ISSN:1583–6258.
6. Smadja NP, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure FX, Birgand G, Holmes AH. Machine learning for clinical decision support in infectious diseases: A narrative review of current applications. Clin Microbiol Infect. 2020;26(5):584–95. https://doi.org/10.1016/j.cmi.2019.09.009.
7. Omar KS, Mondal P, Khan NS, Rizvi RK, Islam N. A machine learning approach to predict autism spectrum disorder. In: The International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, Cox'sBazar, Bangladesh. 2019. https://doi.org/10.1109/ECACE.2019.8679454.
8. Raghavendra S, Santosh KJ. Performance evaluation of random forest with feature selection methods in prediction of diabetes. Int J Elect Comput Eng (IJECE). 2019;10(1):353–9. https://doi.org/10.11591/ijece.v10i1.pp353-359.
9. Wissel T, Pfeiffer T, Frysch R, Knight RT, Chang EF, Hinrichs H, Rieger JW, Rose G. Hidden Markov model and support vector machine based decoding of finger movements using electrocorticography. J Neural Eng. 2013;10(5): 056020. https://doi.org/10.1088/1741-2560/10/5/056020.
10. Mathur P, Kakwani K, Diplav, Kudavelly S, Ramaraju GA. Deep learning based quantification of ovary and follicles using 3D transvaginal ultrasound in assisted reproduction. In: The 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada. 2020. https://doi.org/10.1109/EMBC44109.2020.9176703.
11. Dewi RM, Adiwijaya, Wisesty UN, Jondri. classification of polycystic ovary based on ultrasound images using competitive neural network. J Phys: Conf Ser. 2018;971:012005. https://doi.org/10.1088/1742-6596/971/1/012005.
12. Polycystic ovary syndrome (PCOS). Kaggle. https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos. Accessed 15 Jun 2021.
13. Mehrotra P, Chatterjee J, Chakraborty C, Ghoshdastidar B, Ghoshdastidar S. Automated screening of polycystic ovary syndrome using machine learning techniques. In: The Annual IEEE India Conference. IEEE, Hyderabad, India. 2011. https://doi.org/10.1109/INDCON.2011.6139331.
14. Index of Geo Dataset. https://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS4nnn/GDS4987/. Accessed 15 Jun 2021.
15. Meena K, Manimekalai M, Rethinavalli S. A novel framework for filtering the PCOS attributes using data mining techniques. Int J Eng Res Technol (IJERT). 2015;4(1):702–6. ISSN: 2278–0181.
16. Meena K, Manimekalai M, Rethinavalli S. Correlation of artificial neural network classification and Nfrs attribute filtering algorithm for PCOS data. IJRET: Int J Res Eng Technol. 2015;4(3):519–24. eISSN: 2319–1163.
17. Balogun JA, Egejuru NC, Idowu PA. Comparative analysis of predictive models for the likelihood of infertility in women using supervised machine learning techniques. Comput Rev J. 2018; 2:313–30. ISSN: 2581–6640.
18. PCOS-Survey/PCOSData. Github. 2017. https://github.com/PCOS-Survey/PCOSData. Accessed 15 Jun 2021.
19. Vikas B, Anuhya BS, Chilla M, Sarangi S. A critical study of Polycystic Ovarian Syndrome (PCOS) classification techniques. IJCEM Int J Comput Eng Manage. 2018;21(4).
20. Denny A, Raj A, Ashok A, Ram MC, George R. i-HOPE: Detection and prediction system for Polycystic Ovary Syndrome (PCOS) using machine learning techniques. In: The Proceeding of Region 10 Conference (TENCON). IEEE, Kochi, India. 2019. https://doi.org/10.1109/TENCON.2019.8929674.
21. Bharati S, Podder P, Mondal MRH. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In: The Proceeding of IEEE Region 10 Symposium (TENSYMP). IEEE, Dhaka, Bangladesh. 2020. https://doi.org/10.1109/TENSYMP50017.2020.9230932.
22. Hassan MM, Mirza T. Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome. Int J Comput App. 2020;175(17). https://doi.org/10.5120/ijca2020920688.
23. Neto C, Silva M, Fernandes M, Ferreira D, Machado J. Prediction models for Polycystic Ovary Syndrome using data mining. In: Antipova T. (eds) Advances in Digital Science. ICADS 2021. Adv Intell Syst Comput. 2021;1352. https://doi.org/10.1007/978-3-030-71782-7_19.
24. Munjal A, Khandia R, Gautam B. A machine learning approach for selection of Polycystic Ovarian Syndrome (PCOS) attributes and comparing different classifier performance with the help of WEKA and PyCaret. Int J Sci Res. 2020;59–63. https://doi.org/10.36106/ijsr.
25. Prapty AS, Shitu TT. An efficient decision tree establishment and performance analysis with different machine learning approaches on Polycystic Ovary Syndrome. In: The 23rd International Conference on Computer and Information Technology (ICCIT). DHAKA, Bangladesh. 2020. https://doi.org/10.1109/ICCIT51783.2020.9392666.
26. Nandipati SCR, Ying CX, Wah KK. Polycystic Ovarian Syndrome (PCOS) classification and feature selection by machine learning techniques. Appl Math Comput Intell. 2020;9:65–74. http://dspace.unimap.edu.my:80/xmlui/handle/123456789/69392.
27. Pushkarini H, Anusuya MA. A prediction model for evaluating the risk of developing PCOS. Int Res J Eng Technol (IRJET). 2020;7(9):1150–6. eISSN: 2395–0056.
28. Tanwani N. Detecting PCOS using machine learning. Int J Modern Trends Eng Sci (IJMTES). 2020;7(1):1–20. ISSN: 2348–3121.
29. Thomas N, Kavitha A. Prediction of polycystic ovarian syndrome with clinical dataset using a novel hybrid data mining classification technique. Int J Adv Res Eng Technol (IJARET). 2020;11(11):1872–81. https://doi.org/10.34218/IJARET.11.11.2020.174.
30. Inan MSK, Ulfath RE, Alam FI, Bappee FK, Hasan R. Improved sampling and feature selection to support extreme gradient boosting for PCOS diagnosis. In: The 11th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, NV, USA. 2021. https://doi.org/10.1109/CCWC51732.2021.9375994.
31. Zhang X, Liang B, Zhang J, Hao X, Xu X, Chang HM, Leung PCK, Tan J. Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening. Mol Cell Endocrinol. 2021;523: 111139. https://doi.org/10.1016/j.mce.2020.111139.
32. Hopfield JJ. Artificial neural networks. IEEE Circuits Devices Mag. 1998;4(5):3–10. https://doi.org/10.1109/101.8118.
33. Haq AU, Li JP, Memon MH, Nazir S, Sun R. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mob Inf Syst. 2018. https://doi.org/10.1155/2018/3860146.
34. Lin SK, Hsiu H, Chen HS, Yang CJ. Classification of patients with alzheimer's disease using the arterial pulse spectrum and a multilayer perceptron analysis. Sci Rep. 2021;11:8882. https://doi.org/10.1038/s41598-021-87903-7.
35. Jusman Y, Indra Z, Salambue R, Kanafiah SNAM, Nurkholid MAF. Comparison of multi layered perceptron and radial basis function classification performance of lung cancer Ddata. J Phys Conf Ser. 2020;1471: 012043. https://doi.org/10.1088/1742-6596/1471/1/012043.

36. Das R, Sengur A. Evaluation of ensemble methods for diagnosing of valvular heart disease. Expert Syst Appl. 2010;37(7):5110–5. https://doi.org/10.1016/j.eswa.2009.12.085.
37. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
38. Arora N, Kaur PD. A Bolasso based consistent feature selection enabled random forest classification algorithm: an application to credit risk assessment. Appl Soft Comput J. 2020;86: 105936. https://doi.org/10.1016/j.asoc.2019.105936.
39. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119–39. https://doi.org/10.1006/jcss.1997.1504.
40. Liu Q, Wang X, Huang X, Yina X. Prediction model of rock mass class using classification and regression tree integrated adaboost algorithm based on TBM driving data. Tunn Undergr Space Technol. 2020;106: 103595. https://doi.org/10.1016/j.tust.2020.103595.
41. Ampomah EK, Qin Z, Nyame G. Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. Information. 2020;11(6):332. https://doi.org/10.3390/info11060332.
42. Velliangiria S, Alagumuthukrishnan S, Joseph SIT. A review of dimensionality reduction techniques for efficient computation. Procedia Comput Sci. 2019;165:104–11. https://doi.org/10.1016/j.procs.2020.01.079.
43. Huang C, Huang X, Fang Y, Xu J, Qu Y, Zhai P, Fan L, Yin H, Xu Y, Li J. Sample imbalance disease classification model based on association rule feature selection. Pattern Recogn Lett. 2020;133:280–6. https://doi.org/10.1016/j.patrec.2020.03.016.
44. Khaire, UM, Dhanalakshmi, R. Stability of feature selection algorithm: A review. J King Saud Univ Comput Inform Sci. 2019. https://doi.org/10.1016/j.jksuci.2019.06.012.
45. Zhou Y, Vales MI, Wang A, Zhang Z. Systematic bias of correlation coefficient may explain negative accuracy of genomic prediction. Brief Bioinform. 2017;18(5):744–53. https://doi.org/10.1093/bib/bbw064.
46. Zhou H, Deng Z, Xia Y, Fu M. A new sampling method in particle filter based on Pearson correlation coefficient. Neurocomputing. 2016;216:208–15. https://doi.org/10.1016/j.neucom.2016.07.036.
47. Zebari RR, Abdulazeez AM, Zeebaree DQ, Zebari DA, Saeed JN. A comprehensive review of dimensionality reduction Techniques for feature selection and feature extraction. J Appl Sci Technol Trends. 2020;1(2):56–70. https://doi.org/10.38094/jastt1224.
48. Panthong R, Srivihok A. Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. Procedia Comput Sci. 2015;72:162–9. https://doi.org/10.1016/j.procs.2015.12.117.
49. Chen CW, Tsai YH, Chang FR, Lin WC. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert Syst. 2020;37: e12553. https://doi.org/10.1111/exsy.12553.
50. Zhou Q, Zhou H, Li T. Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. Knowl-Based Syst. 2016;95:1–11. https://doi.org/10.1016/j.knosys.2015.11.010.
51. Panda D, Ray R, Abdullah AA, Dash SR. Predictive systems: Role of feature selection in prediction of heart disease. J Phys: Conf Ser. 2019;1372: 012074. https://doi.org/10.1088/1742-6596/1372/1/012074.
52. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manage Process (IJDKP). 2015;5(2). https://doi.org/10.5281/zenodo.3557376.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.