




# Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models

I. S. Silva<sup>1</sup> · C. N. Ferreira<sup>2</sup> · L. B. X. Costa<sup>3</sup> · M. O. Sôter<sup>4</sup> · L. M. L. Carvalho<sup>3</sup> · J. de C. Albuquerque<sup>4</sup> · M. F. Sales<sup>3</sup> · A. L. Candido<sup>5</sup> · F. M. Reis<sup>6</sup> · A. A. Veloso<sup>1</sup> · K. B. Gomes<sup>3,4</sup> 

Received: 9 June 2021 / Accepted: 1 September 2021 / Published online: 15 September 2021  
© Italian Society of Endocrinology (SIE) 2021

## Abstract

**Purpose** Polycystic Ovary Syndrome (PCOS) is the most frequent endocrinopathy in women of reproductive age. Machine learning (ML) is the area of artificial intelligence with a focus on predictive computing algorithms. We aimed to define the most relevant clinical and laboratory variables related to PCOS diagnosis, and to stratify patients into different phenotypic groups (clusters) using ML algorithms.

**Methods** Variables from a database comparing 72 patients with PCOS and 73 healthy women were included. The BorutaShap method, followed by the Random Forest algorithm, was applied to prediction and clustering of PCOS.

**Results** Among the 58 variables investigated, the algorithm selected in decreasing order of importance: lipid accumulation product (LAP); abdominal circumference; thrombin activatable fibrinolysis inhibitor (TAFI) levels; body mass index (BMI); C-reactive protein (CRP), high-density lipoprotein cholesterol (HDL-c), follicle-stimulating hormone (FSH) and insulin levels; HOMA-IR value; age; prolactin, 17-OH progesterone and triglycerides levels; and family history of diabetes mellitus in first-degree relative as the variables associated to PCOS diagnosis. The combined use of these variables by the algorithm showed an accuracy of 86% and area under the ROC curve of 97%. Next, PCOS patients were gathered into two clusters in the first, the patients had higher BMI, abdominal circumference, LAP and HOMA-IR index, as well as CRP and insulin levels compared to the other cluster.

**Conclusion** The developed algorithm could be applied to select more important clinical and biochemical variables related to PCOS and to classify into phenotypically different clusters. These results could guide more personalized and effective approaches to the treatment of PCOS.

**Keywords** Polycystic Ovary Syndrome · Machine learning · Phenotype

## Introduction

Polycystic Ovary Syndrome (PCOS), a highly inherited complex polygenic, multifactorial disorder, is characterized by anovulatory cycles, clinical and/or biochemical

A. A. Veloso and K. B. Gomes contributed equally to this study.

✉ K. B. Gomes  
karinabgb@gmail.com

<sup>1</sup> Departamento das Ciências da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

<sup>2</sup> Colégio Técnico, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

<sup>3</sup> Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

<sup>4</sup> Departamento de Análises Clínicas e Toxicológicas, Faculdade de Farmácia, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG 31270-901, Brazil

<sup>5</sup> Departamento de Clínica Médica, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

<sup>6</sup> Departamento de Ginecologia e Obstetrícia, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

hyperandrogenism and polycystic ovaries, resulting from endocrine dysfunction that affects women of childbearing age [1, 2].

The prevalence of PCOS varies according to the diagnostic criteria, ranging from 5 to 10%, if androgen excess and anovulation are mandatory manifestations, to as much as 20% according to the Rotterdam consensus that included polycystic ovarian morphology in the trio of core features but defined PCOS as the presence of any two out of the three criteria [2, 3].

PCOS patients present hormonal imbalance, resulting in the formation of cysts from arrested antral follicles, menstrual irregularity and amenorrhea [4]. In addition, there are changes in inflammatory parameters and oxidative stress, contributing to the development of cardiovascular disease (CVD), type 2 diabetes mellitus [DM2] and infertility [5, 6].

The different PCOS phenotypes are based on the presence of hyperandrogenism, oligo-anovulation, and/or polycystic ovarian morphology, according to their different combinations [7]. However, other clinical and laboratory parameters commonly evaluated in women with PCOS are not included in the phenotypical characterization of the patients. A retrospective birth cohort study carried out in 728 women showed that 68% of women who met the diagnostic criteria for PCOS remained undiagnosed [8]. Consequently, several women who are at high risk for the development of CVD, diabetes mellitus, and metabolic disorders, such as obesity, are not diagnosed [1].

In the last years, our group has developed several studies involving biomarkers related to PCOS, to better characterize its pathophysiology and diagnosis. We showed that genetic polymorphisms of cytokine genes may contribute to common metabolic disorders associated with PCOS, and that the disease presents an imbalance between pro- and anti-inflammatory cytokines, with prominent counter-regulatory cytokine production [9–11]. Besides, oxidative stress [12], lipid metabolism [13], hemostatic parameters [14, 15], and other metabolic pathways [16–19] are impaired in PCOS women.

Machine Learning (ML) is the scientific discipline that centers on how computers learn from data, which comprises the intersection of statistics and computer science, focusing on predictive and descriptive computing algorithms [20]. ML includes several analytical algorithms and the main categories of these algorithms are supervised and unsupervised. In supervised learning, an algorithm is trained with a set of input features (or variables) and the outcome of interest (i.e., classes) (e.g., automatic classification), in which the outcomes are used to ‘train’ the algorithm to reach a high predictive power within an acceptable range of accuracy. In unsupervised learning, the outcome of interest is undefined and an algorithm tries to isolate naturally occurring patterns or groups (e.g., clustering analysis) [21].

The potential for improving prediction and visualization quality in research has driven the use of ML in several areas and medical specialties such as therapeutics, oncology, pathology, clinical chemistry and personalized medicine. Currently, large-scale applications and integration with general clinical practice is already a reality [22].

In this study, to define the most relevant biomarkers related to the diagnosis of the PCOS, we evaluated clinical and laboratory variables associated to the pathophysiology of PCOS using the ML algorithms Random Forest [23] and BorutaShap [24]. We also assessed the efficacy of the Random Forest algorithm to automatic classification of PCOS patients in clusters according to the same clinical pattern. This approach may be used as a guideline for future investigations related to disease pathogenesis, and to personalize the better treatment for each patient.

## Materials and methods

### Database

We used the data collected in our own cohort, which included 72 PCOS patients and 73 healthy women as control group. The PCOS group was selected in the Division of Endocrinology of an academic hospital in Belo Horizonte, Brazil, and the control group was recruited among employees and students from the same university in the period from 2011 to 2013. The Ethics Committee (COEP) of the Federal University of Minas Gerais (UFMG) approved the study (CAAE 0379.0.203.000-11). Written informed consent was obtained from all participants.

In those studies, PCOS diagnosis was performed according to European Society of Human Reproduction/Embryology and the American Society for Reproductive Medicine criteria (ESHRE/ASRM) [25], considering the presence of at least two of three criteria: (1) oligo/amenorrhea and anovulation; (2) clinical or laboratory hyperandrogenism and (3) ultrasonography presenting micropolycystic ovaries (presence of 12 or more follicles in the ovary each measuring 2–9 mm in diameter and/or increased ovarian volume > 10 mL).

Exclusion criteria for both groups were current users of steroidal and non-steroidal anti-inflammatory medications, anabolic steroids, isotretinoin, cyclosporine, antiretroviral, insulin and oral contraceptives; presence of diabetes mellitus, kidney and liver disease, thyroid and adrenal disorders, hyperprolactinemia, autoimmune disease, cancer, acute inflammatory disease, orthopedic implant, hypogonadism, and pregnancy; as well as C-reactive protein (CRP) > 10 mg/dL. Patients that presented elevated basal 17-OH progesterone levels were submitted to acute adrenocorticotrophic

hormone (ACTH) stimulation testing to distinguish nonclassical adrenal hyperplasia from PCOS.

Venous blood samples were obtained after 12 h fasting. The aliquots were immediately processed and stored at  $-80^{\circ}\text{C}$  until the analysis. The serum samples were used for measure fasting glucose, C-reactive protein (CRP) and lipid profile using Vitros kits (Johnson and Johnson®). Hormones were also measured in serum samples using Abbott Architect® (chemiluminescent immunoassay), including total testosterone for diagnosis. Microparticles and cytokine levels were evaluated using flow cytometry assay. Thrombin activatable fibrinolysis inhibitor (TAFI), vascular cell adhesion molecule-1 (VCAM-1), proprotein convertase subtilisin/kexin type 9 (PCSK9), haptoglobin and D-dimer levels were measured using ELISA method. Polymorphisms were evaluated by Polymerase Chain Reaction (PCR) allele specific or TaqMan®. All procedures were conducted according to the manufacturer's instructions [9–16].

Hypertension was defined as systolic blood pressure  $>140$  mmHg and/or diastolic blood pressure  $>90$  mmHg at the time of interview, or use of antihypertensive medication. Dyslipidemia was inferred in patients using lipid-lowering medication or diagnosed according to the III Brazilian Guidelines on Dyslipidemia and Atherosclerosis Prevention [26]. The Homeostatic Model Assessment (HOMA) for insulin resistance (IR) was calculated by:  $(\text{insulin (mU/L)} \times \text{glucose (mM/L)})/22.5$  [27]. The Lipid Accumulation Product (LAP) index was determined by:  $((\text{waist circumference cm} - 58) \times (\text{triglycerides mg/dL}))$  [28].

## Data analysis

We included all the clinical and laboratory variables collected in our preceding studies [9–16] in the analysis, which were comprised by genetic polymorphisms, concentrations of biomarkers related to inflammation, lipid and glucose profile, endocrine and hemostatic variables, as well as some index related to metabolism, comorbidities, family history, and life habits (Table 1).

## Machine-learning algorithms

First, we used the BorutaShap [29] feature selection method to select the most important PCOS dataset variables to identify women with PCOS. This process is fundamental to improve the performance and the explicability of the classification model. BorutaShap is an implementation of the Boruta algorithm [30], which iteratively removes the variables, and are less relevant than added artificial variables with random values. However, the BorutaShap measures the variables relevance with the SHapley Additive exPlanations (SHAP) metric [24]. The use of the SHAP metric improves

the original Boruta in speed and quality of the produced variables subset.

Then, using the PCOS dataset and the selected subset of variables, we evaluated the Random Forests Classifier (RFC) in the automatic classification tasks of women in the PCOS and control classes. We selected the RFC because it is very popular and efficient in the classification of tabular (or structured) data [23, 24]. The RFC works as an ensemble of decision trees, in which for the model training, the RFC randomly selects (with replacement) a pre-defined number of subsets of the training data, each subset may be composed of different variables and records (i.e., patients). Next, the RFC trains a decision tree for each subset and the final class of a test record is defined by a majority voting. This behavior reduces bias and variance in the RFC results [23, 24].

Supplemental material Figure 1 shows a decision tree example. A decision tree defines the outcome (i.e., class) moving from the root to a leaf in a structure, in which each node represents a criterion that needs to be attended to move to the next node. In our experiments, we used the Python programming language and scikit-learn Random Forest implementation (<https://scikit-learn.org/stable/>). We used the following default scikit-learn hyperparameters: 100 subsets of the training data (or classifiers in the ensemble), and the number of variables in each subset equals to 9 (i.e.,  $\sqrt{\text{number of variables in dataset}}$ ).

We used the tenfold cross-validation technique to assess the RFC's efficacy in the PCOS dataset, using the BorutaShap selected variables. The tenfold cross-validation technique randomly divides the dataset into tenfold with no replacement. Then, it trains a classifier with the data of ninefold and tests the classifier in onefold not used to the training. This process repeats until each of these tenfold is used to test the model. The final score is the mean of results of all the tenfold. As evaluation metrics, we used the sensitivity, specificity, and area under the receiver-operating characteristics (ROC) curve (AUC).

Next, we used the TreeExplainer [24] approach to analyze the importance of each variable to the created Random Forest classification model (i.e., the importance of each variable to indicate that one patient has PCOS). TreeExplainer is a method that computes optimal local explanations of a tree-based classification model, as defined by the classical game theory Shapley values [24]. The output of the TreeExplainer is called of SHapley Additive exPlanations (SHAP) values.

Finally, we analyzed PCOS patients' clusters using the k-Means algorithm [31] and we evaluated the generated clusters using as input variables according to the SHAP values generated by the TreeExplainer in the explanation of the Random Forest model. These variables were presented as absolute values (mean and standard deviation). We used this approach because the SHAP values indicate the importance of each variable to the model to classify PCOS women [24].

**Table 1** Variables included in the machine-learning algorithms

Group	Variable
Polymorphisms (genotypes frequency]	TGFβ1 (rs 1800470, rs 1800471) IFNγ (rs 2430561) IL6 (rs 1800795) IL10 (rs 1800896, rs 1800871, rs 1800872) PCSK9 (rs562556, rs505151, rs11206510) Haptoglobin (Hp1/Hp2) TNF (rs 1800629) PAI-1 (rs1799889, rs2227631) TAFI (rs3742264, rs1926447, rs940)
Concentrations	MP-cell derived from leucocyte (CD45+) (MP/μL) MP-cell derived from tissue factor (CD142+) (MP/μL) MP-cell derived from platelets (CD41+) (MP/μL) MP-cell derived from endothelium (CD51+) (MP/μL) VCAM -1 (ng/mL) Triglycerides (mg/dL) HDL (mg/dL) LDL (mg/dL) VLDL (mg/dL) Total cholesterol (mg/dL) Fasting glucose (mg/dL) PCSK9 (ng/mL) TAFI (μg/mL) CRP (mg/dL) 17-OH progesterone (ng/dL) 25-hydroxy D vitamin (ng/mL) Haptoglobin (g/L) Insulin (μUI/mL) Prolactin (ng/mL) D-Dimer (ng/mL) FSH (mUI/mL) IL6 (fg/mL) IL10 (fg/mL) TGF (fg/mL)
Index	Age (years) HOMA-IR BMI (kg/m <sup>2</sup> ) Abdominal circumference (cm) LAP
Comorbidities (yes/no]	Dyslipidemia Hypertension Thrombosis Diabetes mellitus
Family history (yes/no]	Dyslipidemia Hypertension Thrombosis Diabetes mellitus PCOS
Habits (yes/no]	Smoker or ex-smoker Alcohol use Sedentary

*TGFβ* transforming growth factor beta; *IFNγ* interferon-gamma; *IL* interleukin; *PCSK9* proprotein convertase subtilisin/kexin type 9; *TNF* tumor necrosis factor; *PAI-1* plasminogen activator inhibitor-1; *TAFI* thrombin activatable fibrinolysis inhibitor; *MP* microparticles; *VCAM1* vascular cell adhesion molecule 1; *HDL* high-density lipoprotein; *LDL* low-density lipoprotein; *VLDL* very low-density lipoprotein; *CRP* C-protein reactive protein; *FSH* follicle-stimulating hormone; *HOMA-IR* Homeostases Model Assessment

**Table 1** (continued)Insulin Resistance; *BMI* body mass index; *LAP* lipid accumulation product; *PCOS* polycystic ovary syndrome

In other words, the SHAP values can indicate the variables that have more probability to predict PCOS cluster for each woman individually.

### Statistical analysis

The normality of the variables was verified by the Shapiro–Wilk test. Categorical variables were described as absolute value and percentage, and continuous variables were expressed as mean  $\pm$  standard deviation or median and interquartile range. Differences in the frequency of variables were assessed by the Chi-square test or Fisher test. The mean difference was assessed by Student *T* test and median difference by Mann–Whitney test. Statistical analyses were conducted using SPSS version 21 software. *p* value  $< 0.05$  was considered significant.

### Results

The PCOS and the control groups had similar age. Body mass index (BMI), abdominal circumference and testosterone levels were higher in PCOS. In addition, more patients with PCOS presented dyslipidemia compared to the control group. Hypertension, sedentary, smokers and alcohol users' frequencies did not differ between the groups (Table 2).

The random forest model showed accuracy = 0.862, sensitivity = 0.869, specificity = 0.857, and precision = 0.880 to predict the PCOS diagnosis. The mean area under the ROC curve obtained by each fold of the cross-validation was  $0.97 \pm 0.03$  (Fig. 1).

The importance of the variable feature within the model was estimated using the SHAP algorithm. The SHAP algorithm showed that, in decreasing order of importance, LAP, abdominal circumference, TAFI levels, body mass index (BMI), CRP, high-density lipoprotein cholesterol (HDL-c), follicle-stimulating hormone (FSH), and insulin levels, HOMA-IR value, age; prolactin, 17-OH progesterone, triglycerides levels; and family history of diabetes mellitus (in first-degree relative) were the variables considered most important for the model to predict PCOS (Fig. 2).

In Fig. 2, according to SHAP, values below 0 on the X-axis, tended to contribute to non-PCOS, while values above 0 tended to contribute to PCOS diagnosis. Red color indicates higher values of the variable and blue color indicates lower values of the variable. Considering the 15 variables more important to the model, higher values of LAP, abdominal circumference, TAFI, BMI, FSH, insulin, HOMA-IR, age, 17-OH-progesterone, and triglycerides; lower HDL-c and prolactin levels; as well as presence of diabetes mellitus in

**Table 2** Characterization of PCOS and control groups

Variables	PCOS <i>N</i> = 72	Control <i>N</i> = 73	<i>p</i>
Age (years)	30.6 $\pm$ 5.2	28.8 $\pm$ 7.2	0.080
BMI (kg/m <sup>2</sup> )	28.4 $\pm$ 8.3	23.1 $\pm$ 5.3	$< 0.001^*$
AC (cm)	96.5 (15.4)	74.0 (15.5)	$< 0.001^*$
Testosterone (mmol/L)	53.30 (48.08)	28.50 (21.35)	$< 0.001^*$
Dyslipidemia			
Yes	9.7%	0%	0.013*
No	90.3%	100%	
Hypertension			
Yes	4.2%	1.4%	0.366
No	95.8%	98.6%	
Sedentary			
Yes	49.0%	50.0%	0.910
No	51.0%	50.0%	
Smoking			
Yes	8.3%	2.7%	0.166
No	91.7%	97.3%	
Alcohol			
Yes	15.3%	23.3%	0.222
No	84.7%	76.7%	

Normal variables are shown as mean  $\pm$  standard deviation. Non-normal variables are shown as median (Interquartile Range). Student *t* test for normal variables and Mann–Whitney test for non-normal variables. Categorical variables were analyzed by Chi-square test and its results are shown as percentage

*BMI* body mass index, *AC* abdominal circumference

Significant\*: *p*  $< 0.05$

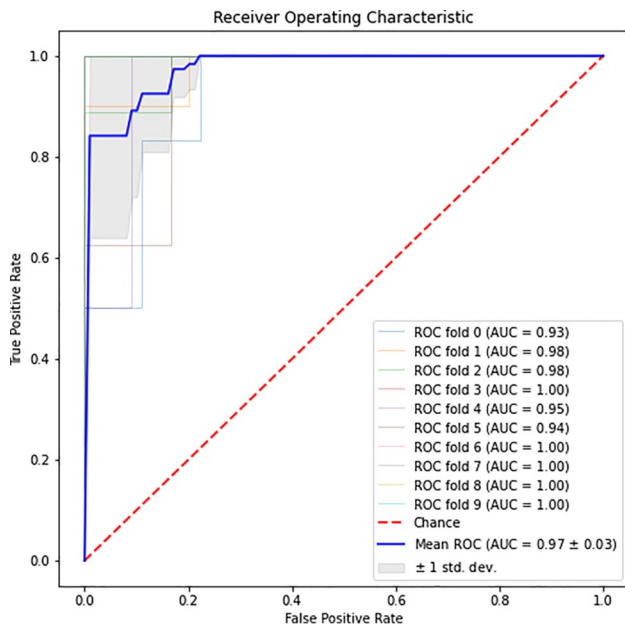
any relative, were related to PCOS diagnosis. Curiously, higher CRP levels were predominantly observed in non-PCOS group.

We applied the same variables to cluster the PCOS patients according to their similarity in SHAP values for each variable. Figure 3 presents the mean absolute values of the variables. Cluster 1 included 54 patients and cluster 2 was comprised of 18 women with PCOS. Comparing the two clusters, BMI ( $p = 1.16 \times 10^{-13}$ ) and abdominal circumference ( $p = 8.23 \times 10^{-9}$ ) showed higher values in cluster 1. Besides, LAP ( $p = 0.001$ ) and HOMA-IR ( $p = 0.005$ ) values, as well as CRP ( $p = 0.012$ ) and insulin levels ( $p = 0.022$ ) were also higher in cluster 1 compared to cluster 2.

### Discussion

The adoption of a mathematical model that allows the diagnosis of PCOS using routine clinical and laboratory variables is highly desirable, especially in those conditions in





**Fig. 1** ROC curves for the prediction of PCOS. The curves represent each cross-validation stage (total tenfold) and the mean curve

which the presence of polycystic ovaries is not observed, and a differential diagnosis with other diseases related to hyperandrogenism is necessary. A model based on routine laboratory tests rather than specialized imaging methods would also allow PCOS screening in primary care settings, thereby increasing the possibility of early lifestyle interventions to prevent complications [32].

This study evaluated routine clinical and laboratory variables as predictors of PCOS using machine-learning algorithms. We observed that the analysis together of 15

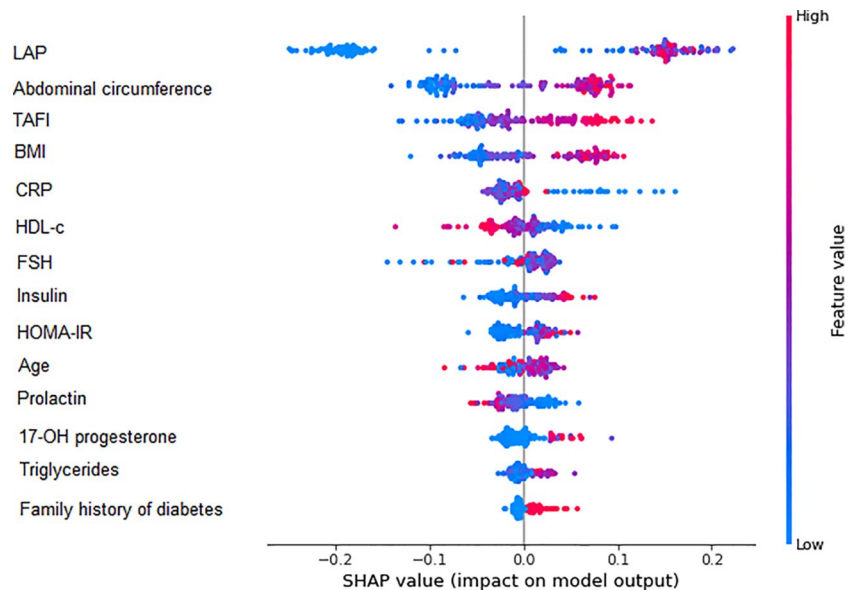
variables through Random Forest model was able to predict PCOS with an accuracy = 0.862. In addition, the application of these variables was able to group the PCOS patients in two clusters with different phenotypes.

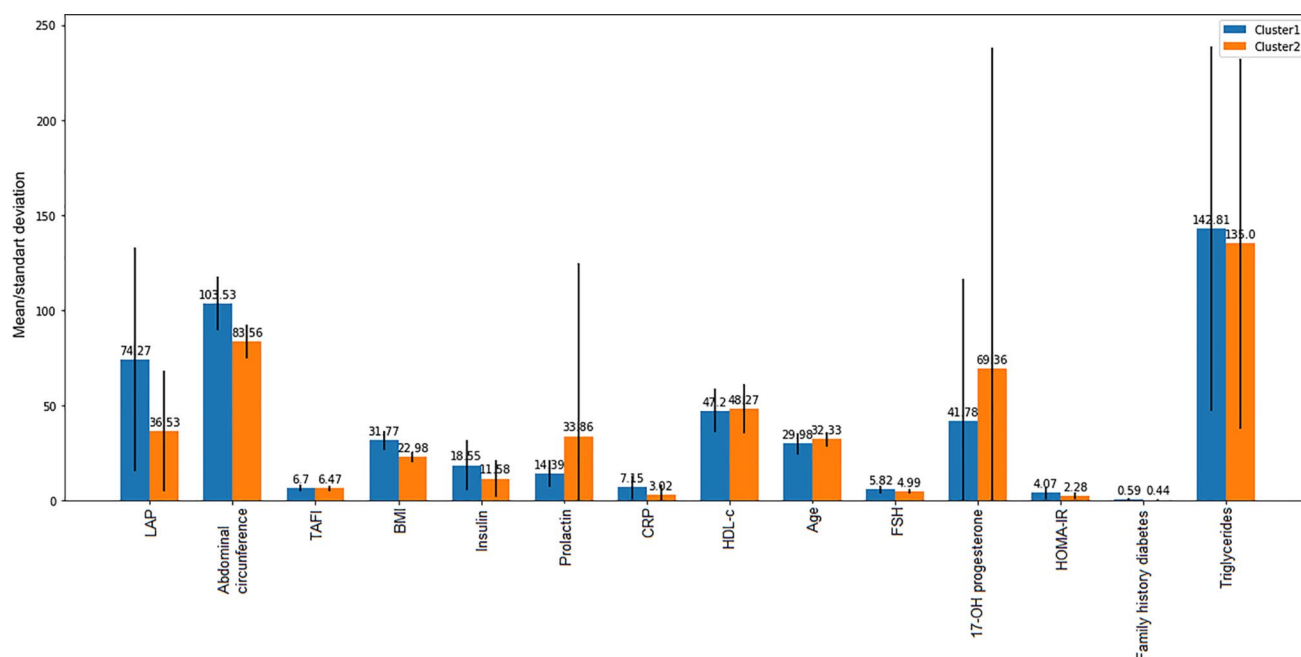
We used the Random Forest because it shows low bias and variance in the results [23, 24]. Besides, Random Forest is very popular and efficient in automatically classifying tabular (or structured) data [23, 24]. These characteristics contribute to the found efficacy in the experiments be close to real life.

According to SHAP, the results confirm our previous findings that showed a trend towards to glucose intolerance, obesity, cardiovascular risk, pro-thrombotic state and dyslipidemia in PCOS individuals [9–16]. However, new results should be highlighted. Higher 17-OH progesterone levels were observed in PCOS group. In fact, Pall et al. (2010) [33] found elevated 17-OH progesterone levels (> 2.0 ng/mL) in 25% and 20% of obese and lean patients with PCOS, respectively. Although serum elevation of 17-OH progesterone is useful as nonclassic adrenal hyperplasia biomarker, higher levels are found in other diseases associated with hirsutism, acne, alopecia, and ovulatory and menstrual dysfunction, such as PCOS [33].

Disturbances in the secretion of the gonadotrophin-releasing hormone (GnRH) results in increased LH: FSH release due to ovarian estrogen that is responsible for causing an abnormal feedback mechanism [34]. Although the isolate measurement of FSH is not a routine practice to PCOS diagnosis, several studies have indicated that FSH gene is one of the susceptible genes for PCOS, which plays a central role in folliculogenesis by stimulating granulosa-cell estrogen production through induction of aromatase activity [35, 36]. Actually, our model showed that higher FSH levels

**Fig. 2** SHAP value for the Random Forest model regarding PCOS prediction. Values below 0 on the X-axis tend to non-PCOS (control) and values above 0 tend to PCOS diagnosis. The red color indicates the greater the value of the variable and the blue color indicates the lower value





**Fig. 3** Variables selected by SHAP model to classify the PCOS patients in two clusters. Variables represented as mean of absolute values

were predominantly present in PCOS women, which suggest that FSH release is important in the pathophysiology of PCOS. In agreement, Deniz et al. [37] also observed significantly higher FSH levels in patients with PCOS compared to controls.

Lower prolactin levels were observed in the PCOS group in our model. This finding is in agreement with other studies that found significantly lower serum prolactin in PCOS patients than in healthy controls [38, 39]. In addition, low prolactin levels within the physiological range have been related to poorer metabolic markers in men [40] and in people aged 80 [41]. A possible explanation to the lower prolactin levels in PCOS is that prolactin is regulated by dopaminergic and serotonergic pathways in the hypothalamus, which also control emotional and mood circuits, both of which tend to be affected by PCOS [38, 40].

In the present study, the mean age of cases and controls was similar but not identical, with the PCOS group being on average 1.8 year older than the control group. This slight difference probably was responsible for the inclusion of age in the statistical model, even though with little discrimination according to the SHAP model. PCOS is a chronic condition often starting in the adolescence and persisting through the reproductive years [2], therefore, a linear association of age with the likelihood of PCOS does not seem plausible.

An important application of our algorithm was the possibility of clustering the PCOS patients according to the similarity of the variables included in the model. According to the international consensus workshop in Rotterdam, four phenotypes of PCOS cases can be identified, based

on PCOS diagnosis criteria, i.e., oligo- and/or anovulation, hyperandrogenism and polycystic ovaries presentation [42]. However, this classification does not take into account the comorbidities or complications of the syndrome, which makes it difficult to adopt more personalized therapeutic procedures. Although the frequency of these PCOS phenotypes has been the aim of extensive research [43–45], clinical outcomes are currently determined according to the Rotterdam phenotypes [46].

Differently, our model classified the PCOS women in two clusters based on similar clinical and biochemical characteristics, considering all the variables significantly associated with PCOS, as defined by the algorithm. Cluster 1 gathered patients with significantly higher BMI, abdominal circumference, LAP and HOMA-IR index, as well as CRP and insulin levels compared to cluster 2. These data suggest that PCOS women can develop different metabolic impairment, characterized by insulin resistance, inflammation, and obesity in one of the phenotypes. Consequently, therapeutic approaches that aim these comorbidities should be applied early in women with this phenotype.

It is important to highlight that CRP values in cluster 2 were very low, which could justify the lower CRP values in the PCOS group when compared to controls in the SHAP model (although within the normal range in both groups). This was unexpected, since PCOS is a sub-chronic inflammatory disease [10], therefore, these findings should be replicated before any generalization based on the characteristics of the syndrome.

The present study has several limitations. First, hyperandrogenism is an alteration commonly seen in both PCOS and nonclassic adrenal hyperplasia, thus distinguishing the two disorders clinically is often difficult. In relation to the small sample size and the cross-sectional design of the study, and to minimize the risk of bias, we chose an algorithm recognized for providing a strong regularization component, thus avoiding overfitting, and the capacity of the model was adjusted for the data available. Besides, its evaluation is robust because it employs cross-validation, which provides statistical guarantees about the error estimate provided. Although an internal validation with accurate case prediction has been implemented, an external validation with another population of the same geographic area, and other populations with different ethnic background should be conducted with our algorithm to assess its potential usefulness in the clinical setting.

## Conclusion

Our study presented an algorithm capable of selecting more important clinical and biochemical variables related to PCOS. Furthermore, the same variables were able to distinguish the PCOS group into two phenotypically different clusters. The application of a validated machine-learning algorithm can be useful to guide more personalized and effective approaches for treating PCOS and preventing of its comorbidities. However, further studies should be conducted to define a reliable diagnostic algorithm applicable in all PCOS cases.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s40618-021-01672-8>.

**Acknowledgements** FMR, AAV and KBG are grateful to CNPq for the research fellowship.

**Author contributions** Conceptualization: ISS, AAV, and KBG. Data curation: CNF, LBXC, MOS, LMLC, JA, MFS, ALC, and FMR. Formal analysis: ISS. Funding acquisition: KBG. Investigation: ISS, AAV, FMR, and KBG. Methodology: ISS and AAV. Project administration: KBG. Resources: KBG. Software: ISS and AAV. Supervision: AAV and KBG. Validation: ISS. Visualization: ISS. Roles/writing—original draft: ISS and KBG. Writing—review and editing: ISS, FMR, and KBG.

**Funding** The grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

**Availability of data and material** The datasets generated during the current study are not publicly available but are available from the corresponding author on reasonable request.

**Code availability** Python programming language and scikit-learn Random Forest implementation (<https://scikit-learn.org/stable/>).

## Declarations

**Conflict of interest** The authors declare that they have no conflicts of interest or financial/personal relationships that could have appeared to influence the work reported in this paper.

**Ethical approval** The Ethics Committee (COEP) of the Federal University of Minas Gerais (UFMG) approved the study (CAAE 0379.0.203.000-11). We certify that the study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki.

**Informed consent** Written informed consent was obtained from all participants.

## References

- Meier R (2018) Polycystic ovary syndrome. *Nurs Clin North Am* 53(3):407–420
- Azziz R (2018) Polycystic ovary syndrome. *Obstet Gynecol* 132(2):321–336
- Bozdag G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO (2016) The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. *Hum Reprod* 31:2841–2855
- Patel S (2018) Polycystic ovary syndrome (PCOS), an inflammatory, systemic, lifestyle endocrinopathy. *J Steroid Biochem Mol Biol* 182:27–36
- Oh J, Lee J, Lee H, Oh Y, Sung Y, Chung H (2009) Serum C-reactive protein levels in normal-weight polycystic ovary syndrome. *Korean J Intern Med* 24(4):350–355
- Hilali N, Vural M, Camuzcuoglu H, Camuzcuoglu A, Nurten A (2013) Increased prolidase activity and oxidative stress in PCOS. *Clin Endocrinol (Oxf)* 79(1):105–110
- National Institutes of Health (2012) Evidence-based methodology workshop on polycystic ovary syndrome. December 3–5. Executive summary. Final report. <https://prevention.nih.gov/docs/programs/pcos/FinalReport.pdf>. Accessed 22 May 2021
- March W, Moore V, Willson K, Phillips D, Norman R, Davies M (2010) The prevalence of polycystic ovary syndrome in a community sample assessed under contrasting diagnostic criteria. *Hum Reprod* 25(2):544–551
- Sóter M, Ferreira C, Sales M, Candido A, Reis FM, Milagres K, Ronda C, Silva I, Sousa M, Gomes K (2015) Peripheral blood-derived cytokine gene polymorphisms and metabolic profile in women with polycystic ovary syndrome. *Cytokine* 76(2):227–235
- Tosatti J, Sóter M, Ferreira C, Silva I, Cândido A, Sousa M, Reis FM, Gomes K (2020) The hallmark of pro- and anti-inflammatory cytokine ratios in women with polycystic ovary syndrome. *Cytokine* 134:155187
- Carvalho L, Ferreira C, Sóter M, Sales M, Rodrigues K, Martins S, Candido A, Reis FM, Silva I, Campos F, Gomes K (2017) Microparticles: inflammatory and haemostatic biomarkers in polycystic ovary syndrome. *Mol Cell Endocrinol* 443:155–162
- Carvalho L, Ferreira C, Oliveira D, Rodrigues K, Duarte R, Teixeira M, Xavier L, Candido A, Reis F, Silva I, Campos F, Gomes K (2017) Haptoglobin levels, but not Hp1-Hp2 polymorphism, are associated with polycystic ovary syndrome. *J Assist Reprod Genet* 34(12):1691–1698
- Xavier L, Sóter M, Sales M, Oliveira D, Reis H, Candido A, Reis FM, Silva I, Gomes K, Ferreira C (2018) Evaluation of PCSK9 levels and its genetic polymorphisms in women with polycystic ovary syndrome. *Gene* 644:129–136



14. Carvalho L, Ferreira C, Candido A, Reis FM, Sôter M, Sales M, Silva I, Nunes F, Gomes K (2017) Metformin reduces total microparticles and microparticles-expressing tissue factor in women with polycystic ovary syndrome. *Arch Gynecol Obstet* 296(4):617–621
15. Sales M, Sôter M, Candido A, Fernandes A, Oliveira F, Ferreira A, Sousa M, Ferreira C, Gomes K (2013) Correlation between plasminogen activator inhibitor-1 (PAI-1) promoter 4G/5G polymorphism and metabolic/proinflammatory factors in polycystic ovary syndrome. *Gynecol Endocrinol* 29(10):936–939
16. Xavier L, Gontijo N, Rodrigues K, Cândido A, Reis F, Sousa M, Silveira J, Oliveira F, Ferreira C, Gomes K (2019) Polymorphisms in vitamin D receptor gene, but not vitamin D levels, are associated with polycystic ovary syndrome in Brazilian women. *J Gynecol Endocrinol* 35(2):146–149
17. Reis G, Gontijo N, Rodrigues K, Alves M, Ferreira C, Gomes K (2017) Vitamin D receptor polymorphisms and the polycystic ovary syndrome: a systematic review. *J Obstet Gynaecol Res* 43(3):436–446
18. Alves M, de Souza I, Ferreira C, Cândido AL, Bizzi M, Oliveira-Reis Gomes FFK (2020) Galectin-3 is a potential biomarker to insulin resistance and obesity in women with polycystic ovary syndrome. *Gynecol Endocrinol* 36(9):760–763
19. Oliveira F, Mamede M, Bizzi M, Rocha A, Ferreira C, Gomes K, Cândido AL, Reis F (2019) Brown adipose tissue activity is reduced in women with polycystic ovary syndrome. *Eur J Endocrinol* 181(5):473–480
20. Rahul C (2015) Machine learning in medicine. *Circulation* 132(20):1920–1930
21. Saber H, Somai M, Rajah G, Scalzo F, Liebeskind D (2019) Predictive analytics and machine learning in stroke and neurovascular medicine. *Neurol Res* 41(8):681–690
22. Handelman G, Kok H, Chandra R, Razavi A, Lee M, Asadi H (2018) eDoctor: machine learning and the future of medicine. *J Intern Med* 284(6):603–619
23. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
24. Lundberg M, Erion G, Chen H, DeGrave A, Prutkin J, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S (2020) From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2(1):56–67
25. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group (2004) The Rotterdam ESHRE/ASRM revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil Steril* 81:19–25
26. Santos R (2001) III Diretrizes Brasileiras Sobre Dislipidemias e Diretriz de Prevenção da Aterosclerose do Departamento de Aterosclerose da Sociedade Brasileira de Cardiologia. *Arq Bras Cardiol* 77(3):1–25
27. Tang Q, Xueqin L, Song P, Xu L (2015) Optimal cut-off values for the homeostasis model assessment of insulin resistance (HOMA-IR) and pre-diabetes screening: Developments in research and prospects for the future. *Drug Discov Ther* 9(6):380–385
28. Lwow F, Jedrzejuk D, Milewicz A, Szmigiero L (2016) Lipid accumulation product (LAP) as a criterion for the identification of the healthy obesity phenotype in postmenopausal women. *Exper Gerontol* 82:81–87
29. Keany E (2021) BorutaShap 1.0.15 2020. <https://pypi.org/project/BorutaShap/>. Accessed 26 May 2021
30. Kursa M, Rudnicki W (2010) Feature selection with the Boruta package. *J Stat Softw* 36(11):1–13
31. Bock H (2007) Clustering methods: a history of k-means algorithms. Selected contributions in data analysis and classification. Springer, Berlin, pp 161–172
32. Teede H, Misso M, Costello M, Dokras A, Laven J, Moran L, Piltonen T, Norman R, International PCOS Network (2018) Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Hum Reprod* 33(9):1602–1618
33. Pall M, Azziz R, Beires J, Pignatelli D (2010) The phenotype of hirsute women: a comparison of polycystic ovary syndrome and 21-hydroxylase-deficient nonclassic adrenal hyperplasia. *Fertil Steril* 94(2):684–689
34. Saadia Z (2020) Follicle stimulating hormone (LH: FSH) ratio in polycystic ovary syndrome (PCOS) obese vs non- obese women. *Med Arch* 74(4):289–293
35. Speiser P, Knochenhauer E, Dewailly D, Fruzzetti F, Marcondes J, Azziz R (2000) A multicenter study of women with nonclassical congenital adrenal hyperplasia: relationship between genotype and phenotype. *Mol Genet Metab* 71:527–534
36. Qiu L, Liu J, Hei Q (2015) Association between two polymorphisms of follicle stimulating hormone receptor gene and susceptibility to polycystic ovary syndrome: a meta-analysis. *Chin Med Sci J* 30(1):44–50
37. Deniz R, Yavuzkir S, Ugur K, Ustebay D, Baykus Y, Ustebay S, Aydin S (2021) Subfatin and asprosin, two new metabolic players of polycystic ovary syndrome. *J Obstet Gynaecol* 41(2):279–284
38. Glinborg D, Altinok M, Mumm H, Buch K, Ravn P, Andersen M (2014) Prolactin is associated with metabolic risk and cortisol in 1007 women with polycystic ovary syndrome. *Hum Reprod* 29:1773–1779
39. Yang H, Di J, Pan J, Yu R, Teng Y, Cai Z, Deng X (2020) The Association between prolactin and metabolic parameters in pcos women: a retrospective analysis. *Front Endocrinol (Lausanne)* 11:263
40. Corona G, Wu F, Rastrelli G, Lee D, Forti G, O'Connor D, O'Neill T, Pendleton N, Bartfai G, Boonen S, Casanueva F, Finn J, Huhtaniemi I, Kula K, Punab M, Vanderschueren D, Rutter M, Maggi M, EMAS Study Group (2014) Low prolactin is associated with sexual dysfunction and psychological or metabolic disturbances in middle-aged and elderly men: the European male aging study (EMAS). *J Sex Med* 11(1):240–253
41. Wagner R, Heni M, Linder K, Ketterer C, Peter A, Bohm A, Hatzigelaki E, Stefan N, Staiger H, Häring H, Fritsche A (2014) Age-dependent association of serum prolactin with glycaemia and insulin sensitivity in humans. *Acta Diabetol* 51(1):71–78
42. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group (2004) Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). *Hum Reprod* 19(1):41–47
43. Cupisti S, Haeberle L, Schell C, Richter H, Schulze C, Hildebrandt T, Oppelt P, Beckmann M, Dittrich R, Mueller A (2011) The different phenotypes of polycystic ovary syndrome: no advantages for identifying women with aggravated insulin resistance or impaired lipids. *Exp Clin Endocrinol Diabetes* 119:502–508
44. Mehrabian F, Khani B, Kelishadi R, Kermani N (2011) The prevalence of metabolic syndrome and insulin resistance according to the phenotypic subgroups of polycystic ovary syndrome in a representative sample of Iranian females. *J Res Med Sci* 16:763–769
45. Shroff R, Syrop C, Davis W, Van Voorhis B, Dokras A (2007) Risk of metabolic complications in the new PCOS phenotypes based on the Rotterdam criteria. *Fertil Steril* 88:1389–1395
46. Lizneva D, Suturina L, Walker W, Brakta S, Gavrilova-Jordan L, Azziz R (2016) Criteria, prevalence, and phenotypes of polycystic ovary syndrome. *Fertil Steril* 106(1):6–15

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