

Original Research

Received 29 June 2026

Revised 28 July 2026

Accepted 11 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (16s): 510–523

KEYWORDS:

Polycystic ovary syndrome; Machine learning; non-invasive screening; Random Forest; SHAP; Calibration; Decision curve analysis; Stepped care.

Explainable Machine Learning for Non-invasive Screening of Polycystic Ovary Syndrome: Development, Internal Validation and Stepped-Care Evaluation.

H Manoj T Gadiyar¹, Dhananjaya G M^{1*}, Shreesha¹, Sinchana S N¹, Sumedha¹, Sujnan Acharya¹

¹Canara Engineering College, Benjanapadavu, Mangalore, Karnataka 574219, Visvesvaraya Technological University, Belagavi, Karnataka 590018, India

ABSTRACT: Polycystic ovary syndrome (PCOS) affects 10–13% of women of reproductive age, yet diagnosis is often delayed because confirmation requires hormonal assays and pelvic ultrasound. We developed and internally validated a machine learning model that estimates PCOS risk from 16 non-invasive history, symptom and examination features, and quantified what hormonal and ultrasound tests add. Using a public multicentre dataset of 541 women (177 with PCOS) from Kerala, India, logistic regression, support vector machine, random forest and XGBoost models were built in leakage-free pipelines, compared by nested cross-validation and evaluated once on a stratified 20% held-out test set with bootstrap confidence intervals, calibration and decision curve analysis (TRIPOD+AI).

The random forest, selected on nested cross-validation (AUC 0.891 ± 0.026), achieved a test AUC of 0.884 (95% CI 0.803–0.951), with sensitivity 77.8% and specificity 90.4% at a threshold fixed on training data; calibration-in-the-large was good (intercept 0.04; slope 1.41, 95% CI 0.97–2.27), and the model gave higher net benefit than referring all or no women. Logistic regression performed comparably (AUC 0.860; DeLong $p = 0.220$). Adding serum hormones gave no significant gain (AUC 0.893), whereas adding ultrasound raised it to 0.945. SHAP analysis identified excess hair growth, skin darkening, weight gain, irregular cycle and menstrual duration as the leading predictors. Non-invasive features can triage women for confirmatory testing; external validation is needed before clinical use.

1. INTRODUCTION

Polycystic ovary syndrome (PCOS) is the most common endocrine disorder of women of reproductive age, with a prevalence of approximately 10–13% depending on the diagnostic criteria applied [1, 2]. It is characterised by hyperandrogenism, ovulatory dysfunction and polycystic ovarian morphology, and is associated with infertility, obesity, insulin resistance, type 2 diabetes and elevated cardiometabolic risk [3–6]. Women with PCOS also experience a markedly higher burden of depressive and anxiety symptoms [7]. In India, prevalence estimates vary widely but indicate a substantial and growing public-health concern [8].

Diagnosis is based on the Rotterdam criteria, as refined by the 2023 International Evidence-based Guideline, and requires two of three features—clinical or biochemical hyperandrogenism, ovulatory dysfunction, and polycystic ovarian morphology on ultrasound or elevated anti-Müllerian hormone (AMH)—after exclusion of mimicking disorders [1, 9]. Because presentation is heterogeneous and no single test is definitive, diagnosis is often delayed: in an international survey, more than one-third of women waited over two years and nearly half consulted three or more health professionals before diagnosis [10], with a

considerable associated economic burden [11]. Where hormonal assays and specialist ultrasound are not routinely available, a low-cost first-tier tool that identifies women who warrant confirmatory assessment could shorten this pathway.

Machine learning (ML) is increasingly applied to clinical prediction [12–14]. For PCOS, studies have examined gene signatures [15], urinary steroid profiles [16] and transcriptomic biomarkers [17], and have classified ovarian ultrasound images with neural networks [18, 19]. Among tabular clinical models, Zad et al. predicted PCOS from electronic health records of 30,601 women (AUC $\approx$ 0.85) [20], Panjwani et al. combined symptom and general-health data in a stacking ensemble (AUC 0.93) [21], and Agirsoy and Oehlschlaeger reported an AUC of 0.985 for XGBoost using clinical and ultrasound features [22]. Other work has explored explainable feature selection [23, 24], ensemble classifiers [25, 26], web-based decision support [27, 28], discriminant analysis [29, 30] and federated learning [31].

Three gaps limit the translation of this literature into early screening. First, most high-performing models rely on hormonal, AMH or ultrasound inputs [22, 23]—the very investigations a pre-referral screen should prioritise—and the incremental value of these tests over history and examination has rarely been quantified. Second, a systematic review found substantial methodological heterogeneity and risk of bias in PCOS ML studies [32]: performance is commonly reported as accuracy from a single split, without confidence intervals, calibration or comparison with logistic regression, and preprocessing before data splitting can leak information and inflate results [33, 34]. Third, few studies follow the TRIPOD+AI reporting guideline [35] or evaluate clinical utility with decision curve analysis [36].

This study addresses these gaps. Its aims were (i) to develop and internally validate an explainable ML model that estimates PCOS risk from non-invasive features only; (ii) to benchmark it rigorously against logistic regression and other learners using nested cross-validation, bootstrap confidence intervals, calibration and decision curve analysis; (iii) to quantify, in a stepped-care framework, how much hormonal and ultrasound testing adds to non-invasive screening, and how much performance depends on predictors that overlap with the diagnostic criteria; and (iv) to translate the model into calibrated risk bands and a deployment pipeline for a web-based screening tool. The tool is intended to prompt timely clinical consultation, not to replace diagnosis.

2. MATERIALS AND METHODS

2.1 Study design and reporting

This was a retrospective diagnostic prediction-model development study with internal validation, reported according to TRIPOD+AI [35]. The workflow and the stepped-care evaluation are summarised in Figure 1. Study design left figure data flow from the source cohort through cleaning, a stratified split into development and held-out test sets, model development and evaluation, and interpretation and deployment. Right: the stepped-care framework, in which non-invasive screening (Tier 1) is compared with the addition of serum hormones (Tier 2) and pelvic ultrasound (Tier 3). BP, blood pressure; CV, cross-validation; DCA, decision curve analysis; LR, logistic regression; RF, random forest; SVM, support vector machine.

2.2 Data source and participants

We used the publicly available PCOS dataset compiled by Kottarathil [37], which contains clinical, anthropometric, biochemical and ultrasound records of 541 women of reproductive age assessed at 10 hospitals in Kerala, India. This dataset has been used in several previous ML studies [22, 23], which facilitates comparison. Each record carries a binary PCOS label (177 PCOS, 364 non-PCOS; prevalence 32.7%) assigned by the treating clinicians; the diagnostic criteria applied by the contributing hospitals are not documented in the source, which we consider in the limitations. The data are fully de-identified.

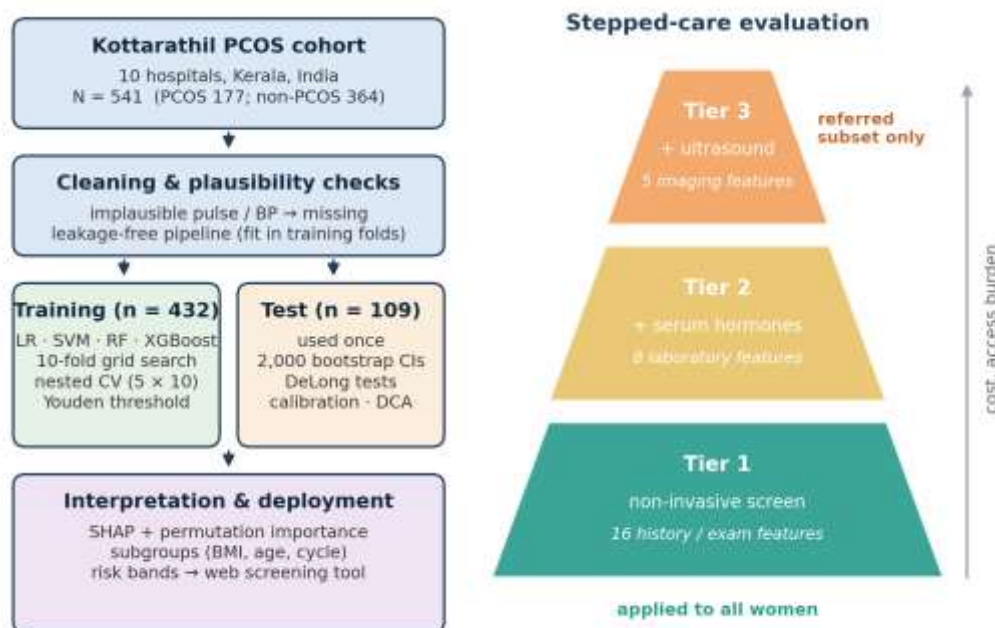


Figure 1. Study design and stepped-care evaluation framework

2.3 Predictors

Predictors were pre-specified into three tiers according to cost and access (Table 1). Tier 1 comprised 16 non-invasive features obtainable from history, a symptom questionnaire and a basic examination: age, body mass index (BMI), waist circumference, waist–hip ratio, menstrual duration, cycle regularity, systolic and diastolic blood pressure, pulse rate, weight gain, excess hair growth, skin darkening, hair loss, acne, fast-food intake and regular exercise. The source variable labelled “cycle length” ranges from 0 to 12 days and therefore records the duration of menstrual bleeding rather than the inter-menstrual interval; it is reported here as menstrual duration. Tier 2 added eight serum measurements (FSH, LH, AMH, TSH, prolactin, vitamin D3, progesterone and random blood sugar), and Tier 3 further added five ultrasound measurements (left and right follicle counts, mean follicle sizes and endometrial thickness). Variables directly related to pregnancy or infertility treatment were excluded.

Table 1. Predictors in each tier of the stepped-care evaluation.

Tier	k	Predictors
Tier 1: non-invasive	16	Age, BMI, waist circumference, waist–hip ratio, menstrual duration, cycle regularity (regular/irregular), systolic and diastolic blood pressure, pulse rate; weight gain, excess hair growth, skin darkening, hair loss, acne, fast-food intake, regular exercise (yes/no)
Tier 2: + hormones	24	Tier 1 + FSH, LH, AMH, TSH, prolactin, vitamin D3, progesterone, random blood sugar
Tier 3: + ultrasound	29	Tier 2 + follicle number (left, right), mean follicle size (left, right), endometrial thickness
Sensitivity analysis	14	Tier 1 without cycle regularity and excess hair growth (components of the diagnostic criteria)

AMH, anti-Müllerian hormone; BMI, body mass index; FSH, follicle-stimulating hormone; LH, luteinising hormone; TSH, thyroid-stimulating hormone.

2.4 Data cleaning and preprocessing

Physiologically implausible values were set to missing before modelling using pre-specified ranges: pulse rate outside 40–150 bpm (2 values), systolic blood pressure outside 70–200 mmHg (1), diastolic blood pressure outside 40–130 mmHg (1) and BMI outside 13–60 kg/m² (1); one invalid cycle-regularity code was also set to missing. The data were then split into a development set (80%, $n = 432$) and a held-out test set (20%, $n = 109$; 36 PCOS) by stratified random sampling. All data-dependent steps—mean imputation of continuous and mode imputation of binary variables, winsorisation of continuous variables at $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, and z-score standardisation—were implemented within a single pipeline whose parameters were learned only from the training folds and applied unchanged to validation and test data, preventing information leakage [33]. No resampling was applied, to preserve calibration.

2.5 Sample size

With 177 events and 16 candidate parameters in Tier 1, the dataset provides 11.1 events per candidate parameter. Using the criteria of Riley et al. [38] with a conservative anticipated Cox–Snell R^2 of 0.108, 1257 participants would be needed to target a global shrinkage factor of 0.9, whereas 339 are needed to estimate overall risk precisely. The available sample therefore meets the latter but not the former criterion; we mitigated overfitting through penalisation, tree-depth constraints, nested cross-validation and explicit calibration assessment, and we treat the sample size as a limitation.

2.6 Model development

Four learners were compared. Penalised (L2) logistic regression served as the reference model, because flexible methods frequently offer no advantage over regression for tabular clinical data [34]. A support vector machine (SVM) with linear or radial basis function kernel [39] produced probabilities by Platt scaling [40]; a random forest (RF) of 300 trees [41]; and gradient-boosted trees (XGBoost) [42] were included as nonlinear learners. Hyperparameters were tuned by grid search with stratified 10-fold cross-validation on the development set, maximising the AUC. To obtain an optimism-reduced estimate and to choose the primary model without using the test set, the complete tuning procedure was repeated within nested cross-validation (5 outer $\times$ 10 inner folds) [43]; the learner with the highest nested AUC was pre-specified as the primary model.

2.7 Performance evaluation

Each final model was refitted on the full development set and evaluated once on the test set. Discrimination was quantified by the AUC and by the area under the precision–recall curve (AUPRC), which is informative under class imbalance [44]. Accuracy, sensitivity, specificity, positive and negative predictive values (PPV, NPV), F1-score and the Matthews correlation coefficient (MCC) were computed at the conventional 0.5 threshold and at an operating threshold that maximised Youden’s index on out-of-fold development predictions, never on the test set. Ninety-five percent confidence intervals (CIs) were obtained from 2,000 bootstrap resamples [45], and paired AUCs were compared with DeLong’s test [46]. Calibration was assessed with calibration plots, the calibration intercept and slope, and the Brier score [47], and clinical utility with decision curve analysis over threshold probabilities of 0.05–0.80 [36].

2.8 Explainability, stepped-care, sensitivity and subgroup analyses

Because impurity-based RF importance is biased towards continuous variables [48], the primary model was explained with SHapley Additive exPlanations (SHAP) [49] and with permutation importance on the test set (mean AUC decrease over 50 permutations). In the stepped-care analysis, the primary learner was retrained and evaluated identically with Tier 2 and Tier 3 predictors. To examine incorporation bias [50], the Tier 1 model was also refitted without cycle regularity and excess hair growth, which form part of the diagnostic criteria. Performance in subgroups defined by BMI (Asian action point of 23 kg/m² [51]), age (< 30 vs ≥ 30 years) and cycle regularity was estimated from nested out-of-fold predictions for all 541 women. Finally, predicted risks were grouped into three bands fixed a priori (low < 0.20 ; moderate 0.20–0.50; high ≥ 0.50), and the observed PCOS proportion in each band was reported.

2.9 Software and deployment

Analyses used Python 3.12.3, scikit-learn 1.8.0 [52], XGBoost 3.4.1, SHAP 0.52.0 and statsmodels, with a fixed random seed (42). The final primary pipeline, including its fitted preprocessing parameters, was serialised for integration into the PCOS Prediction & Care System web application, whose architecture is shown in Figure 7.

3. RESULTS

3.1 Participant characteristics

Of the 541 women, 177 (32.7%) had PCOS. Compared with women without PCOS, those with PCOS were younger, had higher BMI and waist circumference and shorter menstrual duration, and more often reported irregular cycles, weight gain, excess hair growth, skin darkening, hair loss, acne and fast-food intake; waist–hip ratio, blood pressure and regular exercise did not differ materially (Table 2). Missing data after cleaning were minimal (at most 2 values for any predictor).

Table 2. Characteristics of participants by PCOS status.

Characteristic	PCOS (n = 177)	Non-PCOS (n = 364)	p
Age, years	29.00 (27.00–33.00)	32.00 (28.00–36.00)	< 0.001
BMI, kg/m ²	25.10 (23.08–28.30)	23.60 (21.36–26.09)	< 0.001
Waist circumference, inch	35.00 (32.00–37.00)	34.00 (32.00–35.00)	< 0.001
Waist–hip ratio	0.90 (0.86–0.93)	0.89 (0.86–0.93)	0.655
Menstrual duration, days	5.00 (3.00–5.00)	5.00 (5.00–6.00)	< 0.001
Systolic BP, mmHg	110.00 (110.00–120.00)	110.00 (110.00–120.00)	0.901
Diastolic BP, mmHg	80.00 (70.00–80.00)	80.00 (70.00–80.00)	0.510
Pulse rate, bpm	72.00 (72.00–74.00)	72.00 (72.00–74.00)	0.006
Irregular cycle, n (%)	94 (53.4%)	56 (15.4%)	< 0.001
Weight gain, n (%)	121 (68.4%)	83 (22.8%)	< 0.001
Excess hair growth, n (%)	101 (57.1%)	47 (12.9%)	< 0.001
Skin darkening, n (%)	110 (62.1%)	56 (15.4%)	< 0.001
Hair loss, n (%)	102 (57.6%)	143 (39.3%)	< 0.001
Acne, n (%)	123 (69.5%)	142 (39.0%)	< 0.001
Fast-food intake, n (%)	139 (78.5%)	139 (38.3%)	< 0.001
Regular exercise, n (%)	51 (28.8%)	83 (22.8%)	0.157

Continuous variables are median (IQR), compared with the Mann–Whitney U test; binary variables are n (%), compared with the χ^2 test. BP, blood pressure.

3.2 Discrimination and classification

Nested cross-validation AUCs were 0.891 ± 0.026 for RF, 0.886 ± 0.024 for XGBoost, 0.881 ± 0.026 for logistic regression and 0.872 ± 0.030 for SVM; RF was therefore the primary model (selected hyperparameters: maximum depth 6, minimum samples per leaf 5, features per split sqrt). On the held-out test set, RF achieved an AUC of 0.884 (0.803–0.951) and an AUPRC of 0.820 (0.694–0.922) (Table 3; Figure 2). XGBoost (0.895), logistic regression (0.860) and SVM (0.861) performed similarly, and none of the differences from RF was statistically significant (DeLong $p = 0.283, 0.220$ and 0.211 , respectively). The close agreement between nested and test-set AUCs indicates limited optimism.

At the default 0.5 threshold, RF showed high specificity (91.8%) but modest sensitivity (72.2%), because predicted risks for a minority class are conservative. At the operating threshold of 0.39 fixed on development data, sensitivity rose to 77.8% (95% CI 63.6%–90.6%) with specificity 90.4% (83.5%–96.3%), PPV 80.0% and NPV 89.2%; of 36 women with PCOS in the test set, 28 were flagged and 8 were missed, while 7 of 73 women without PCOS were flagged for further testing.

Table 3. Test-set performance of the four models (n = 109; 36 PCOS).

Metric	Logistic regression	SVM	Random forest	XGBoost
AUC	0.860 (0.768–0.936)	0.861 (0.770–0.937)	0.884 (0.803–0.951)	0.895 (0.825–0.952)
AUPRC	0.784 (0.642–0.909)	0.790 (0.645–0.916)	0.820 (0.694–0.922)	0.817 (0.685–0.923)
Accuracy	0.862 (0.798–0.927)	0.862 (0.798–0.927)	0.853 (0.780–0.917)	0.844 (0.771–0.908)
Sensitivity	0.750 (0.600–0.886)	0.750 (0.600–0.892)	0.722 (0.559–0.871)	0.694 (0.533–0.840)
Specificity	0.918 (0.849–0.973)	0.918 (0.848–0.973)	0.918 (0.849–0.973)	0.918 (0.848–0.973)
PPV	0.818 (0.680–0.943)	0.818 (0.680–0.938)	0.812 (0.667–0.939)	0.806 (0.657–0.931)
NPV	0.882 (0.806–0.949)	0.882 (0.803–0.949)	0.870 (0.791–0.943)	0.859 (0.778–0.931)
F1-score	0.783 (0.667–0.879)	0.783 (0.667–0.882)	0.765 (0.638–0.870)	0.746 (0.613–0.853)
MCC	0.684 (0.532–0.821)	0.684 (0.530–0.823)	0.661 (0.502–0.809)	0.638 (0.472–0.785)
Brier score	0.122 (0.080–0.170)	0.120 (0.081–0.166)	0.122 (0.089–0.160)	0.118 (0.082–0.160)
Calibration intercept	-0.01 (-0.63–0.57)	-0.03 (-0.57–0.47)	0.04 (-0.41–0.47)	0.10 (-0.42–0.59)
Calibration slope	0.85 (0.57–1.34)	0.95 (0.65–1.52)	1.41 (0.97–2.27)	1.07 (0.77–1.63)
Nested-CV AUC, mean $\pm$ SD	0.881 $\pm$ 0.026	0.872 $\pm$ 0.030	0.891 $\pm$ 0.026	0.886 $\pm$ 0.024
DeLong p vs RF	0.220	0.211	—	0.283

Values are point estimates (95% bootstrap CI, 2,000 resamples). Threshold-dependent metrics use a probability threshold of 0.5. AUPRC, area under the precision–recall curve; CV, cross-validation; MCC, Matthew’s correlation coefficient; NPV/PPV, negative/positive predictive value.

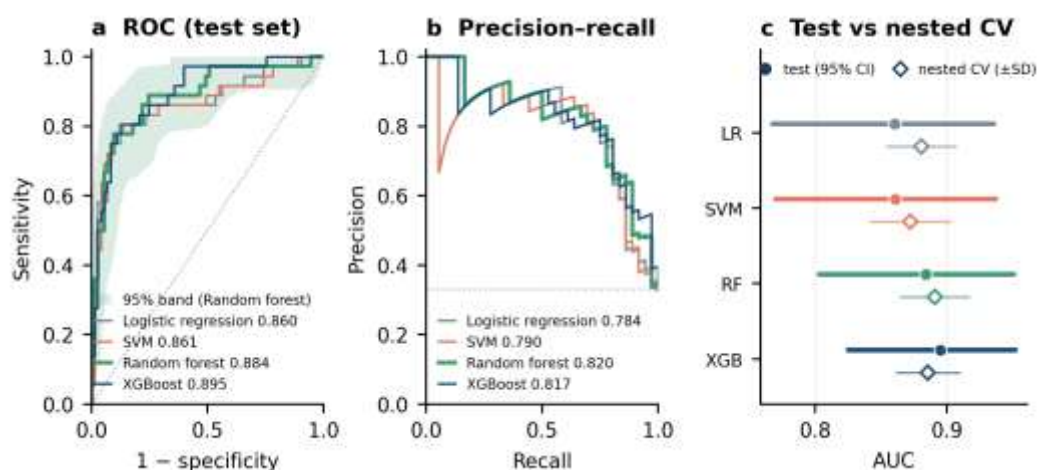


Figure 2. Discrimination on the held-out test set. (a) Receiver operating characteristic curves with the 95% bootstrap band for the primary random forest model; legend values are AUCs. (b) Precision–recall curves; the dotted line marks outcome prevalence. (c) Test-set AUC with 95% CI (circles) compared with nested cross-validation AUC $\pm$ SD (diamonds) for each learner.

3.3 Calibration and clinical utility

RF showed good calibration-in-the-large, with a calibration intercept of 0.04 (-0.41–0.47), a slope of 1.41 (0.97–2.27) and a Brier score of 0.122 (Figure 3a); a slope above 1 indicates that its predictions were somewhat conservative (closer to the average risk than warranted), whereas XGBoost (1.07), SVM (0.95) and logistic regression (0.85) had slopes closer to 1. In decision curve analysis, RF provided a higher net benefit than referring all or no women across threshold probabilities from 0.07 to 0.80 (Figure 3b), the range in which a clinician would reasonably decide on referral for hormonal and ultrasound assessment.

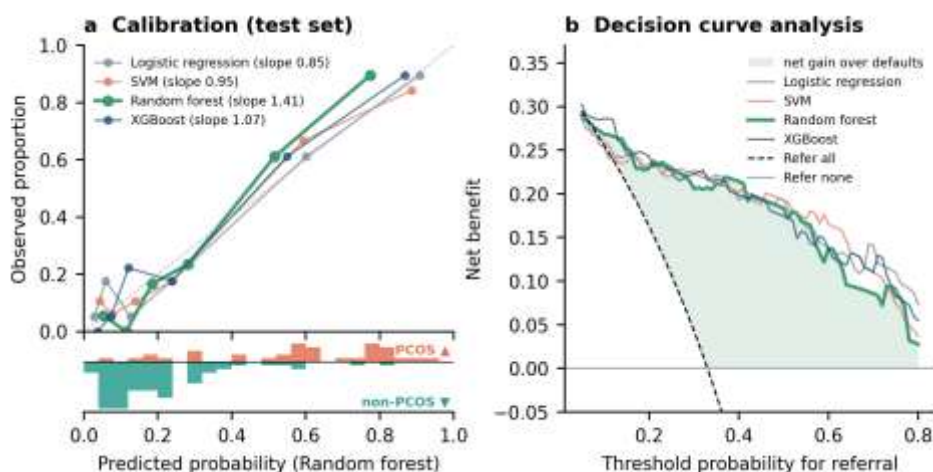


Figure 3. (a) Calibration of the four models on the test set (quantile bins; legend shows calibration slopes), with the mirrored histogram of random forest predictions for women with (top) and without (bottom) PCOS. (b) Decision curve analysis; the shaded area shows the net benefit gained by random-forest-guided referral over the better of the refer-all and refer-none strategies.

3.4 Explainability

In SHAP analysis, excess hair growth, skin darkening, weight gain, irregular cycle and menstrual duration contributed most to predicted risk (Figure 4a). Positive responses for the symptom items and an irregular cycle increased predicted risk, as did higher BMI and waist circumference, whereas a longer menstrual duration lowered it, consistent with the oligomenorrhoea and hyperandrogenic phenotype of PCOS. Permutation importance gave a similar ranking, with skin darkening (0.038 mean AUC decrease), excess hair growth (0.026) and irregular cycle (0.023) as the most influential features on the test set (Figure 4b).

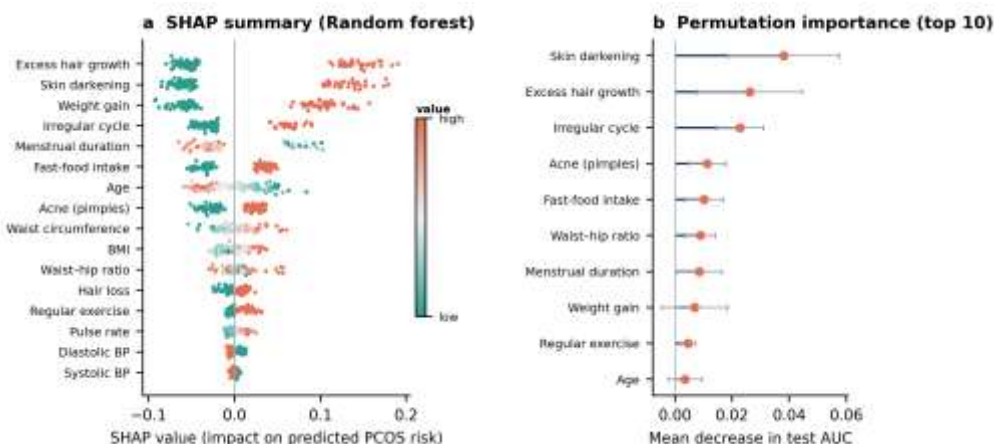


Figure 4. Explainability of the random forest model. (a) SHAP summary: each point is one test-set participant; its horizontal position shows the effect of the feature on predicted PCOS risk and its colour the feature value. (b) Permutation importance (mean decrease in test-set AUC over 50 permutations $\pm$ SD) for the ten most important features.

3.5 Stepped-care, sensitivity and subgroup analyses

Adding serum hormones changed the test AUC only marginally, from 0.884 to 0.893 (DeLong $p = 0.444$), whereas further adding ultrasound increased it significantly to 0.945 ($p = 0.025$; Table 4; Figure 5a), mirroring the diagnostic weight of polycystic ovarian morphology in the guideline [1]. Removing cycle regularity and excess hair growth reduced the Tier 1 AUC modestly, from 0.884 to 0.859 (nested AUC 0.863; $p = 0.128$), indicating that discrimination is not driven solely by predictors that overlap with the diagnostic criteria. Notably, the Tier 2 model without ultrasound did not outperform Tier 1, suggesting that in this cohort the hormonal panel carries little information beyond history and examination.

Across subgroups, nested out-of-fold AUCs ranged from 0.84 to 0.91, with overlapping confidence intervals (Figure 5b). Performance was similar in women with BMI below and above 23 kg/m² (0.91 vs 0.87) and in younger and older women (0.90 vs 0.88), and lower within strata of cycle regularity (0.86 regular; 0.84 irregular), where one of the strongest predictors is held constant.

Table 4. Stepped-care evaluation and sensitivity analysis (random forest).

Model	k	Test AUC (95% CI)	Sens.	Spec.	Nested-CV AUC	p vs Tier 1
Tier 1: non-invasive	16	0.884 (0.804–0.950)	72.2%	91.8%	0.891 ± 0.026	—
Tier 1 without cycle regularity and hair growth	14	0.859 (0.776–0.931)	58.3%	94.5%	0.863 ± 0.030	0.128
Tier 2: + hormones	24	0.893 (0.824–0.950)	61.1%	91.8%	0.888 ± 0.028	0.444
Tier 3: + hormones + ultrasound	29	0.945 (0.885–0.987)	80.6%	98.6%	0.955 ± 0.022	0.025

Sens., sensitivity; Spec., specificity (threshold 0.5). p values from DeLong's test on the test set. k, number of predictors.

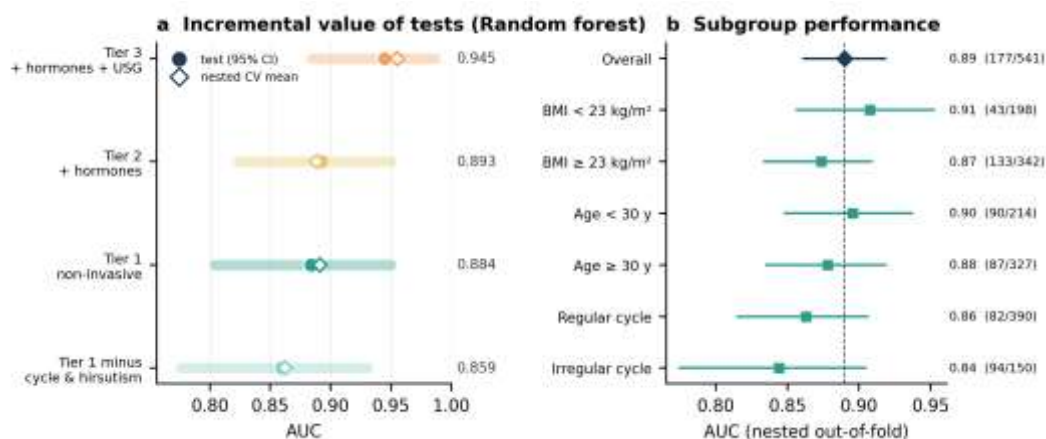


Figure 5. (a) Incremental value of hormonal and ultrasound testing over non-invasive screening, and the effect of removing criterion-overlapping predictors: test-set AUC with 95% CI (circles and bars) and nested cross-validation mean (diamonds). (b) Subgroup performance from nested out-of-fold predictions for all 541 women (AUC with 95% CI; events/n shown on the right).

3.6 Risk stratification and deployment

Using nested out-of-fold predictions, 253 women (46.8%) were classified as low risk, of whom 6.3% had PCOS; 142 (26.2%) as moderate risk (31.0% PCOS); and 146 (27.0%) as high risk (80.1% PCOS) (Figure 6). The observed PCOS proportion thus rose steeply across bands, supporting their use to prioritise confirmatory testing. The deployment pipeline (Figure 7) applies the frozen preprocessing and model to 16 user-entered items, returns the risk band with the individual's leading SHAP contributors and general advice, and displays a disclaimer that the output is a screening estimate.

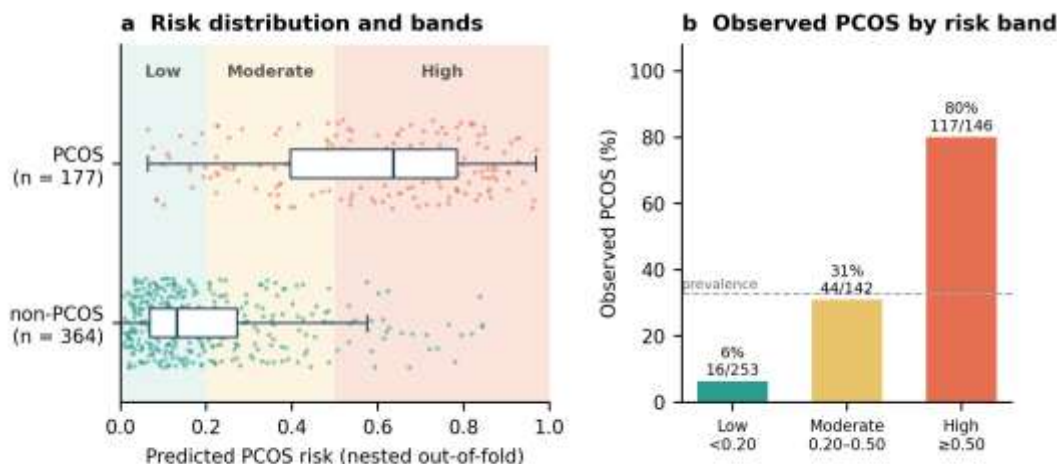


Figure 6. Risk stratification by the random forest model (nested out-of-fold predictions, $n = 541$). (a) Distribution of predicted risk in women with and without PCOS, with the three pre-specified risk bands. (b) Observed PCOS proportion in each band (number with PCOS/number in band); the dashed line marks overall prevalence.

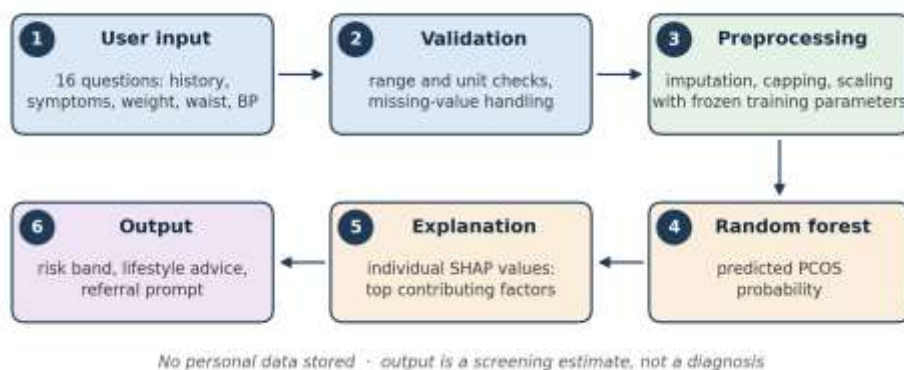


Figure 7. Deployment architecture of the PCOS Prediction & Care System. User inputs are range-checked and processed with the preprocessing parameters learned during development; the model returns a probability, which is mapped to a risk band and accompanied by the individual's leading SHAP contributors and advice. SHAP, SHapley Additive exPlanations.

4. DISCUSSION

4.1 Principal findings

Using 16 non-invasive features from a multicentre Indian cohort, an RF model discriminated women with and without PCOS with a test AUC of 0.884 and a nested cross-validation AUC of 0.891, showed good calibration-in-the-large with modestly conservative predictions, and provided clinical net benefit across plausible referral thresholds. At an operating threshold fixed on development data it identified 77.8% of women with PCOS. Logistic regression, SVM and XGBoost performed comparably. Serum hormones added little beyond non-invasive features, whereas ultrasound raised the AUC to 0.945; removal of criterion-overlapping items caused only a modest, non-significant loss.

4.2 Comparison with previous studies

Previous studies on the same dataset have reported accuracies above 90% and AUCs up to 0.985 [22, 23], but these models incorporate AMH, hormonal and ultrasound variables. Our Tier 3 model, which includes these inputs, reached an AUC of 0.945, approaching the values reported with full clinical data under a stricter, test-set-blind protocol, whereas the non-invasive Tier 1 model provides a realistic estimate of what can be achieved before any test is ordered (Table 5). Our non-invasive

performance is similar to that of a large electronic-health-record model [20] and a symptom-based ensemble [21]. The absence of a meaningful advantage of flexible learners over logistic regression agrees with a systematic review of clinical prediction models [34]; where performance is similar, a simpler model may be preferred for transparency [53].

Table 5. Comparison with selected machine learning studies of PCOS prediction.

Study	Data	Inputs	Best model	Reported performance
Zad et al., 2024 [20]	EHR, 30,601 women (USA)	Clinical + laboratory	Gradient-boosted trees	AUC $\approx$ 0.85
Panjwani et al., 2025 [21]	Two merged public datasets	Symptoms + general health	Stacking ensemble	Accuracy 92.8%; AUC 0.93
Agirsoy & Oehlschlaeger, 2025 [22]	Kottarathil dataset, 541 women	Clinical + ultrasound	XGBoost	Accuracy 93.8%; AUC 0.985
Present study, Tier 3	Kottarathil dataset, 541 women	Non-invasive + hormones + ultrasound	Random forest	AUC 0.945
Present study, Tier 1	Kottarathil dataset, 541 women	16 non-invasive features	Random forest	AUC 0.884; nested 0.891

EHR, electronic health records. Values are as reported by the original authors; studies differ in design and validation, so figures are not directly comparable.

4.3 Clinical and public-health implications

The stepped-care results suggest a practical pathway. A questionnaire-and-examination screen can place most women into low- or high-risk bands; in this cohort only 6.3% of women in the low band had PCOS, whereas 80.1% of those in the high band did. Directing hormonal and ultrasound testing towards moderate- and high-risk women could therefore reduce unnecessary testing while prioritising those most likely to benefit, which is particularly relevant where laboratory and imaging capacity is limited. The prominence of anthropometric and symptom features is consistent with the high prevalence of overall and central obesity in PCOS [54], and the modifiable lifestyle items offer an opportunity to deliver guideline-concordant advice [1]. Deployment should follow principles of responsible ML, including transparent communication of uncertainty and monitoring for performance drift [55][56].

4.4 Strengths and limitations

Strengths include a pre-specified, leakage-free pipeline; selection of the primary model on nested cross-validation rather than the test set; comparison with a logistic regression baseline; bootstrap confidence intervals, calibration and decision curve analysis; SHAP-based explanation; a stepped-care analysis that quantifies the value of invasive tests; a sensitivity analysis for incorporation bias; subgroup evaluation; and use of a public dataset that allows full reproducibility.

Several limitations apply. First, the dataset is modest (541 women) and falls below the sample size recommended to limit overfitting [38], and the test set is small, so confidence intervals are wide. Second, all data come from one Indian state, and the model has not been externally validated; performance in other populations and in primary-care settings, where PCOS prevalence is lower than the 32.7% observed here, is unknown, and predictive values would change accordingly. Third, the diagnostic criteria used to label PCOS in the source hospitals are not documented, and the reference standard may have incorporated some predictors [50]; our sensitivity analysis suggests that this effect is modest. Fourth, symptom and lifestyle items were recorded as simple yes/no responses and may be subject to reporting bias, and variables such as psychological stress, sleep and family history, which may be informative, were unavailable. Finally, the source variable for cycle length records menstrual duration, so the inter-menstrual interval, a key feature of ovulatory dysfunction, could not be modelled.

4.5 Future work

Future work will validate and, if necessary, recalibrate the model prospectively in consecutive women presenting to primary care and gynaecology clinics in different regions; collect the inter-menstrual interval, a structured hirsutism score, family history, stress and sleep; evaluate a formal two-stage referral pathway in which Tier 1 screening precedes hormonal and ultrasound testing; and assess the effect of the web tool on time to diagnosis. Privacy-preserving federated training across institutions [31] will be explored to enlarge the development sample.

5. CONCLUSION

An explainable random forest model using 16 non-invasive features estimated PCOS risk with good discrimination (test AUC 0.884), acceptable calibration and clinical net benefit on internal validation, and stratified women into risk bands with markedly different PCOS prevalence. Ultrasound, but not serum hormones, added substantial discrimination, supporting a stepped-care strategy in which low-cost screening prioritises women for confirmatory assessment. External validation is required before the web-based tool is used in clinical practice.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Department of Information Science and Engineering, and Department of Artificial Intelligence and Machine Learning, Canara Engineering College, Benjanapadavu, for providing the necessary technical support and research facilities to carry out this work. The authors also extend their heartfelt thanks to Canara Engineering College (Autonomous), Mangaluru, for providing the facilities and academic environment essential for carrying out this study. The authors gratefully acknowledge the Visvesvaraya Technological University (VTU), Belagavi, for offering a platform that fosters research and innovation among students and faculty members.

AUTHOR CONTRIBUTIONS

H Manoj T Gadiyar: Conceptualization, Supervision, Domain Expertise, Writing—Review & Editing.

Dhananjaya G M: Conceptualization, Investigation, Supervision, Domain Expertise, Writing—Review & Editing.

Ramya N.R.: Clinical Validation, Writing—Review & Editing.

Shreesha: Data Curation, Software, Formal Analysis.

Sinchana S N: Methodology, Software, Visualization.

Sumedha: Data Curation, Investigation, Writing—Original Draft.

Sujnan Acharya: Software, Validation, Writing—Original Draft.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

This study was a secondary analysis of a publicly available, fully de-identified dataset. No new data were collected and no participants were contacted; institutional ethical approval and informed consent were therefore not required.

DATA AVAILABILITY STATEMENT

The dataset analysed in this study is publicly available on Kaggle (<https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos>). The analysis code and trained model are available from the corresponding author upon reasonable request.

REFERENCES

- [1] Teede, H.J., Tay, C.T., Laven, J.J.E., Dokras, A., Moran, L.J., Piltonen, T.T., et al., 2023. Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Journal of Clinical Endocrinology & Metabolism* 108(10), 2447–2469. <https://doi.org/10.1210/clinem/dgad463>

- [2] Bozdag, G., Mumusoglu, S., Zengin, D., Karabulut, E., Yildiz, B.O., 2016. The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. *Human Reproduction* 31(12), 2841–2855. <https://doi.org/10.1093/humrep/dew218>
- [3] Stener-Victorin, E., Teede, H., Norman, R.J., Legro, R., Goodarzi, M.O., Dokras, A., et al., 2024. Polycystic ovary syndrome. *Nature Reviews Disease Primers* 10(1), 27. <https://doi.org/10.1038/s41572-024-00511-3>
- [4] Joham, A.E., Norman, R.J., Stener-Victorin, E., et al., 2022. Polycystic ovary syndrome. *The Lancet Diabetes & Endocrinology* 10(9), 668–680. [https://doi.org/10.1016/S2213-8587\(22\)00163-2](https://doi.org/10.1016/S2213-8587(22)00163-2)
- [5] Sachdeva, G., Gainder, S., Suri, V., Sachdeva, N., Chopra, S., 2019. Obese and non-obese polycystic ovarian syndrome: comparison of clinical, metabolic, hormonal parameters, and their differential response to clomiphene. *Indian Journal of Endocrinology and Metabolism* 23(2), 257–262.
- [6] Codner, E., 2025. The impact of type 1 diabetes on polycystic ovary syndrome development. *Journal of Pediatric Endocrinology & Metabolism* 38(4), 345–352.
- [7] Cooney, L.G., Lee, I., Sammel, M.D., Dokras, A., 2017. High prevalence of moderate and severe depressive and anxiety symptoms in polycystic ovary syndrome: a systematic review and meta-analysis. *Human Reproduction* 32(5), 1075–1091. <https://doi.org/10.1093/humrep/dex044>
- [8] Ganie, M.A., Vasudevan, V., Wani, I.A., Baba, M.S., Arif, T., Rashid, A., 2019. Epidemiology, pathogenesis, genetics & management of polycystic ovary syndrome in India. *Indian Journal of Medical Research* 150(4), 333–344.
- [9] 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. *Fertility and Sterility* 81(1), 19–25. <https://doi.org/10.1016/j.fertnstert.2003.10.004>
- [10] Gibson-Helm, M., Teede, H., Dunaif, A., Dokras, A., 2017. Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome. *Journal of Clinical Endocrinology & Metabolism* 102(2), 604–612. <https://doi.org/10.1210/jc.2016-2963>
- [11] Azziz, R., Marin, C., Hoq, L., Badamgarav, E., Song, P., 2005. Health care-related economic burden of the polycystic ovary syndrome during the reproductive life span. *Journal of Clinical Endocrinology & Metabolism* 90(8), 4650–4658. <https://doi.org/10.1210/jc.2005-0628>
- [12] Rajkomar, A., Dean, J., Kohane, I., 2019. Machine learning in medicine. *New England Journal of Medicine* 380(14), 1347–1358. <https://doi.org/10.1056/NEJMr1814259>
- [13] Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- [14] Kapadia, D., Jain, R., 2023. PCOS prediction: advancements in medical informatics using artificial intelligence. In: *2023 2nd International Conference on Futuristic Technologies (INCOFT)*. IEEE, pp. 1–5.
- [15] Zhang, X.Z., Pang, Y.L., Wang, X., Li, Y.H., 2018. Computational characterization and identification of human polycystic ovary syndrome genes. *Scientific Reports* 8(1), 12949.
- [16] Dhayat, N.A., Marti, N., Kollmann, Z., Troendle, A., Bally, L., Escher, G., et al., 2018. Urinary steroid profiling in women hints at a diagnostic signature of the polycystic ovary syndrome: a pilot study considering neglected steroid metabolites. *PLoS One* 13(10), e0203903. <https://doi.org/10.1371/journal.pone.0203903>
- [17] Chen, W., Miao, J., Chen, J., Chen, J., 2025. Development of machine learning models for diagnostic biomarker identification and immune cell infiltration analysis in PCOS. *Journal of Ovarian Research* 18.
- [18] Dewi, R.M., Adiwijaya, Wisesty, U.N., Jondri, 2018. Classification of polycystic ovary based on ultrasound images using competitive neural network. *Journal of Physics: Conference Series* 971, 012005.
- [19] Divekar, A., Sonawane, A., 2025. Leveraging AI for automatic classification of PCOS using ultrasound imaging. *arXiv preprint arXiv:2501.01984*.
- [20] Zad, Z., Jiang, V.S., Wolf, A.T., Wang, T., Cheng, J.J., Paschalidis, I.C., et al., 2024. Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. *Frontiers in Endocrinology* 15, 1298628. <https://doi.org/10.3389/fendo.2024.1298628>
- [21] Panjwani, B., Yadav, J., Mohan, V., Agarwal, N., Agarwal, S., 2025. Optimized machine learning for the early detection of polycystic ovary syndrome in women. *Sensors* 25(4), 1166. <https://doi.org/10.3390/s25041166>
- [22] Agirsoy, M., Oehlschlaeger, M.A., 2025. A machine learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features. *Scientific Reports* 15, 33638. <https://doi.org/10.1038/s41598-025-10453-9>

- [23] Elmannai, H., El-Rashidy, N., Mashal, I., Alohal, M.A., Farag, S., El-Sappagh, S., et al., 2023. Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence. *Diagnostics* 13(8), 1506. <https://doi.org/10.3390/diagnostics13081506>
- [24] Khanna, V.V., Chadaga, K., Sampathila, N., Prabhu, S., Bhandage, V., Hegde, G.K., 2023. A distinctive explainable machine learning framework for detection of polycystic ovary syndrome. *Applied System Innovation* 6(2), 32. <https://doi.org/10.3390/asi6020032>
- [25] Mohi Uddin, K.M., et al., 2025. Early PCOS detection: a comparative analysis of machine learning models. *Engineering Reports* 7(2), e70008.
- [26] Faris, N., Sahi, A., Diykh, M., Abdulla, S., Siuly, S., 2025. Enhanced polycystic ovary syndrome diagnosis model leveraging a K-means based genetic algorithm and ensemble approach. *Intelligence-Based Medicine* 11, 100253. <https://doi.org/10.1016/j.ibmed.2025.100253>
- [27] Tiwari, S., Kane, L., Koundal, D., Jain, A., Alhudhaif, A., Polat, K., et al., 2022. SPOSDS: a smart polycystic ovary syndrome diagnostic system using machine learning. *Expert Systems with Applications* 203, 117592. <https://doi.org/10.1016/j.eswa.2022.117592>
- [28] Rahman, M.M., et al., 2024. Empowering early detection: a web-based machine learning approach for PCOS prediction. *Informatics in Medicine Unlocked* 47.
- [29] Joshi, R., et al., 2023. Deep linear discriminant analysis with variation for polycystic ovary syndrome classification. *arXiv preprint arXiv:2303.14401*.
- [30] Gupta, A., et al., 2022. Discriminant analysis in contrasting dimensions for polycystic ovary syndrome prognostication. *arXiv preprint arXiv:2201.03029*.
- [31] Morris, L., et al., 2023. Federated learning on patient data for privacy-protecting polycystic ovary syndrome treatment. *arXiv preprint arXiv:2308.11220*.
- [32] Barrera, F.J., Brown, E.D.L., Rojo, A., Obeso, J., Plata, H., Lincango, E.P., et al., 2023. Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review. *Frontiers in Endocrinology* 14, 1106625. <https://doi.org/10.3389/fendo.2023.1106625>
- [33] Kaufman, S., Rosset, S., Perlich, C., Stitelman, O., 2012. Leakage in data mining: formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data* 6(4), 15. <https://doi.org/10.1145/2382577.2382579>
- [34] Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., Van Calster, B., 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* 110, 12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>
- [35] Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., et al., 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [36] Vickers, A.J., Elkin, E.B., 2006. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 26(6), 565–574. <https://doi.org/10.1177/0272989X06295361>
- [37] Kottarathil, P., 2020. Polycystic ovary syndrome (PCOS) [dataset]. Kaggle. <https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos>
- [38] Riley, R.D., Ensor, J., Snell, K.I.E., Harrell, F.E., Martin, G.P., Reitsma, J.B., et al., 2020. Calculating the sample size required for developing a clinical prediction model. *BMJ* 368, m441. <https://doi.org/10.1136/bmj.m441>
- [39] Cortes, C., Vapnik, V., 1995. Support-vector networks. *Machine Learning* 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- [40] Platt, J.C., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, A.J., Bartlett, P., Schölkopf, B., Schuurmans, D. (Eds.), *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, pp. 61–74.
- [41] Breiman, L., 2001. Random forests. *Machine Learning* 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [42] Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [43] Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7, 91. <https://doi.org/10.1186/1471-2105-7-91>

- [44] Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- [45] Efron, B., Tibshirani, R.J., 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- [46] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3), 837–845.
- [47] Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W., 2019. Calibration: the Achilles heel of predictive analytics. *BMC Medicine* 17(1), 230. <https://doi.org/10.1186/s12916-019-1466-7>
- [48] Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 8, 25. <https://doi.org/10.1186/1471-2105-8-25>
- [49] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Curran Associates, Red Hook, NY, pp. 4765–4774.
- [50] Worster, A., Carpenter, C., 2008. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias. *Canadian Journal of Emergency Medicine* 10(2), 174–175.
- [51] WHO Expert Consultation, 2004. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *The Lancet* 363(9403), 157–163. [https://doi.org/10.1016/S0140-6736\(03\)15268-3](https://doi.org/10.1016/S0140-6736(03)15268-3)
- [52] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.
- [53] Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- [54] Lim, S.S., Davies, M.J., Norman, R.J., Moran, L.J., 2012. Overweight, obesity and central obesity in women with polycystic ovary syndrome: a systematic review and meta-analysis. *Human Reproduction Update* 18(6), 618–637. <https://doi.org/10.1093/humupd/dms030>
- [55] Rathod, V. N., Goudar, R. H., Patil, M., Hukkeri, G. S., & Kaliwal, R. B. (2024, December). Enhancing autism spectrum disorder identification: a machine learning approach using CatBoost. In *2024 International Conference on Innovation and Novelty in Engineering and Technology (INNOVA)* (Vol. 1, pp. 1-5). IEEE.
- [56] Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., et al., 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine* 25(9), 1337–1340. <https://doi.org/10.1038/s41591-019-0548-6>