e-ISSN: 2986-9838

Sriwijaya Journal of Obstetrics and Gynecology (SJOG)

Journal website: https://phlox.or.id/index.php/sjog

Development and Validation of an Explainable Machine-Learning Model to Predict Cumulative Live Birth in PCOS-Associated Infertility

Anies Fatmawati[1*], Hesti Putri[2], Theresia Putri Sinaga[3]

[1]Department of Obstetrics and Gynecology, Halmahera Community Health Center, Halmahera, Indonesia

[2]Department of Informatics Science, CMHC Research Center, Palembang, Indonesia

[3]Department of Public Health, CMHC Research Center, Palembang, Indonesia

ARTICLE INFO

Article history:
Received: August 1, 2025
Accepted: September 20, 2025
Published: October 30, 2025

Keywords:
Forecasting
In vitro fertilization
Live birth rate
Machine learning
Polycystic ovary syndrome

***Corresponding author:**
Anies Fatmawati

E-mail address:
anies.fatmawati@enigma.or.id

All authors have reviewed and approved the final version of the manuscript.

https://doi.org/10.59345/sjog.v3i2.282

A B S T R A C T

Introduction: Polycystic ovary syndrome (PCOS) is the leading cause of anovulatory infertility, yet predicting the cumulative live birth rate (CLBR) after in vitro fertilization (IVF) is difficult because of marked clinical heterogeneity. Conventional linear models discriminate modestly and most machine-learning (ML) tools remain uninterpretable. We aimed to develop and internally validate an explainable ML model for CLBR in women with PCOS.

Methods: In a multicenter retrospective cohort at two tertiary reproductive-medicine centers in Indonesia (January 2018–December 2023), 1,245 women with PCOS (Rotterdam criteria) undergoing a first IVF/intracytoplasmic sperm injection cycle were analysed. Five algorithms (logistic regression, support vector machine, random forest, gradient boosting and eXtreme Gradient Boosting [XGBoost]) were trained (70%) and internally validated (30%) with 5-fold cross-validation. Discrimination used the area under the ROC curve (AUC) with DeLong confidence intervals (CI); SHapley Additive exPlanations (SHAP) quantified feature importance. Multivariable logistic regression, calibration and number-needed-to-treat (NNT) were also derived.

Results: Overall CLBR was 54.6% (95% CI 51.8–57.4%). XGBoost performed best (AUC 0.82, 95% CI 0.79–0.85; accuracy 76.4%; sensitivity 78.2%; specificity 74.1%) and significantly exceeded logistic regression (AUC 0.72; ΔAUC 0.10, $p<0.001$). SHAP ranked top-quality embryo number, maternal age, anti-Müllerian hormone and oocyte yield as dominant predictors. Each additional top-quality embryo raised CLBR odds (adjusted OR 1.85, 95% CI 1.66–2.06, $p<0.001$); ≥3 versus <3 embryos yielded an NNT of 3.3.

Conclusion: An explainable XGBoost model accurately predicts CLBR in PCOS-associated infertility and can support individualised counselling and cycle-management decisions. Prospective external validation is warranted before clinical deployment.

1. Introduction

Polycystic ovary syndrome (PCOS) is the most common endocrinopathy of reproductive-aged women and the principal cause of anovulatory subfertility, affecting an estimated 8–13% of women and reaching prevalences approaching 20% when the broader Rotterdam criteria are applied[1]. Across South and Southeast Asia, including Indonesia, the rising demand for assisted reproductive technology (ART) means that women with PCOS now constitute a substantial proportion of the infertile population presenting to tertiary reproductive-medicine centers[2]. For these patients and their clinicians, an early, individualised estimate of the probability of a live birth is among the most valued pieces of

information, shaping expectations, treatment intensity and financial commitment.

PCOS is defined by the Rotterdam combination of oligo-anovulation, clinical or biochemical hyperandrogenism and polycystic ovarian morphology, and the 2023 international guideline reaffirmed this framework while emphasising structured infertility care[1]. The syndrome's reproductive phenotype is shaped by insulin resistance, hyperandrogenism and dysregulated gonadotropin secretion[1,3,4]. During ART this produces a paradox: an expanded antral follicle pool and high anti-Müllerian hormone (AMH) generate abundant oocytes, yet impaired oocyte competence, asynchronous folliculogenesis and a heightened risk of ovarian hyperstimulation syndrome (OHSS) temper the translation of oocyte quantity into live birth[5,6]. Obesity, common in PCOS, further depresses oocyte and endometrial quality and compounds obstetric risk, so that the path from ovarian reserve to a healthy baby is distinctly non-linear[7,8].

The cumulative live birth rate (CLBR)—at least one live birth resulting from a single initiated ART cycle, encompassing the fresh transfer and all subsequent frozen–thawed embryo transfers—is the patient-centerd efficacy metric now widely recommended as the patient-centerd primary outcome for ART research[6,9]. Reliable, individualised CLBR prediction allows clinicians to counsel realistically, to triage candidates for additional cycles, and to tailor strategies such as elective freeze-all or pre-treatment weight optimisation[10,11].

Despite decades of effort, prognostic models for IVF remain imperfect. A systematic review of IVF prediction models found that most are methodologically weak, poorly calibrated and seldom externally validated, and that few are specific to defined phenotypes such as PCOS[12,13]. Conventional logistic regression assumes additive, linear effects and therefore struggles with the complex interactions that characterise the PCOS cycle—where, for example, the benefit of additional oocytes plateaus and may reverse at high yields, and where the freeze-all strategy itself improves live birth and reduces OHSS, thereby altering the very outcome being predicted[14].

Machine learning (ML) offers a route to capture such non-linear, high-order interactions. Across the ART pathway, ML and artificial-intelligence (AI) methods now inform oocyte and embryo assessment, ovarian-stimulation management and outcome prediction, and head-to-head comparisons suggest that tree-ensemble algorithms can match or exceed classical regression for IVF outcomes[12,15,16]. Yet two barriers have slowed clinical adoption: the opacity of complex 'black-box' models, which undermines clinician trust and shared decision-making, and inconsistent, often inadequate reporting[17,18].

We therefore sought to develop and internally validate an explainable ML model for CLBR in women with PCOS, using routinely available baseline and cycle-specific variables from two Indonesian tertiary centers. We benchmarked five algorithms—including XGBoost, a scalable gradient-boosted tree method—against logistic regression, rendered the best model interpretable with SHapley Additive exPlanations (SHAP), and reported the work in line with contemporary prediction-model standards[13,17-19]. To our knowledge this is among the first multicenter, PCOS-specific, explainable CLBR models derived from Southeast Asian data.

The promise of AI in reproductive medicine has at times outpaced rigorous evidence, with many published models limited by single-center derivation, absent calibration, opaque reporting and a lack of phenotype specificity[13,15]. A credible contribution must therefore do more than report a high AUC: it should be derived from multiple centers, benchmarked against an interpretable standard, calibrated, explained at the level of the individual patient, and reported transparently. We designed the present study to meet each of these expectations within the specific and clinically demanding context of PCOS-associated infertility.

2. Methods

Study design and setting

This was a multicenter retrospective cohort study and clinical prediction-model development with internal validation, conducted at two tertiary referral reproductive-medicine centers in Indonesia between January 2018 and December 2023. The study and its reporting followed the STROBE recommendations for observational research and the TRIPOD+AI statement for machine-learning prediction models. The study was approved by CMHC Research Center, Indonesia (approval No. CMHC/EC/2023/0214); because data were retrospective and fully de-identified, the requirement for individual informed consent was waived.

Participants

Eligible participants were women aged 20–40 years with a confirmed diagnosis of PCOS by the revised 2003 Rotterdam criteria (≥2 of oligo-anovulation, hyperandrogenism, polycystic ovarian morphology) undergoing their first stimulated IVF or intracytoplasmic sperm injection (ICSI) cycle[1]. Exclusion criteria were severe male-factor infertility (azoospermia), preimplantation genetic testing cycles, untreated hydrosalpinx, uterine anomalies, and incomplete records (>10% of variables missing).

Outcome

The primary outcome was cumulative live birth, defined as at least one live birth (delivery of a viable infant at ≥24 weeks' gestation) arising from one initiated ART cycle, including the fresh transfer and all subsequent frozen–thawed embryo transfers (FET) originating from that cycle[9]. Gestational age was established from the last menstrual period and confirmed by first-trimester Hadlock ultrasound biometry. Secondary perinatal outcomes—mode of delivery (classified by ACOG indication), gestational age at birth (weeks+days), birth weight (grams), Apgar scores at 1 and 5 minutes, neonatal intensive-care-unit (NICU) admission, hypertensive disorders of pregnancy, gestational diabetes and moderate/severe OHSS (Golan classification)—were abstracted among women achieving live birth and compared by transfer strategy (fresh versus freeze-all/FET); singleton and multiple gestations were distinguished. Because oligomenorrhoea is near-universal in PCOS, gestational age was assigned from the date of oocyte retrieval and cross-checked against first-trimester Hadlock biometry, rather than relying on the last menstrual period. Moderate/severe OHSS was graded by the Golan classification, and the ≥24-week viability threshold was applied identically at both centers. Each woman contributed exactly one (first) cycle, so observations were independent. Women were followed for a minimum of 12 months after the index retrieval, and those with cryopreserved embryos remaining unused at the data cut were classified as not having achieved cumulative live birth from transferred embryos, a conservative definition.

Variables and data collection

Predictor variables were extracted from electronic medical records. Baseline features comprised maternal age, body mass index (BMI, kg/m^2), duration and type of infertility, and the endocrine profile (baseline FSH, LH, estradiol, AMH [ng/mL] and antral follicle count [AFC]). Cycle features comprised the controlled ovarian stimulation (COS) protocol (GnRH antagonist versus agonist), total gonadotropin dose, duration of stimulation and trigger type. Embryological features comprised the number of oocytes retrieved, maturation and fertilization rates, and the number of top-quality embryos on day 3 or day 5. A top-quality embryo was defined a priori by consensus morphological grading (good-grade cleavage-stage embryos with ≥7 even blastomeres and <10% fragmentation on day 3, or blastocysts of Gardner grade ≥3BB on day 5), applied uniformly across centers after grader harmonisation. AMH was measured on the same automated immunoassay platform at both centers to limit inter-assay variability. Parity and prior pregnancy history were recorded and considered as candidate predictors. Blood pressure was recorded with the participant seated, using Korotkoff phase V; haematological and endocrine assays used each center's accredited laboratory reference ranges.

Sample size

All consecutive eligible cycles in the study window were included. For a model incorporating up to ten candidate predictors, the achieved sample of 1,245 women with 680 events provided 68 events per variable—far exceeding the conventional minimum of 10–20 events per variable—supporting stable estimation and limiting overfitting; all inference used a two-sided α of 0.05.

Machine-learning pipeline

Continuous variables were standardised and missing values imputed with k-nearest-neighbours imputation. Data were partitioned into a training set (70%) and an internal validation set (30%) with stratification on the outcome. Five algorithms were developed and compared: logistic regression (the interpretable baseline), support vector machine, random forest, gradient boosting and XGBoost[13]. Hyperparameters were tuned by grid search with 5-fold cross-validation on the training set to mitigate overfitting; the validation set was held out for unbiased performance estimation. Analyses used Python 3.9 with scikit-learn and XGBoost.

Statistical analysis

Baseline characteristics were compared using the Mann–Whitney U test for continuous variables and the chi-square test for categorical variables, with effect sizes reported as Cohen's d or odds ratios (OR) with 95% confidence intervals (CI). Model discrimination was quantified by the area under the ROC curve (AUC) with DeLong 95% CIs, and pairwise AUC differences tested by the DeLong method; calibration was assessed by the calibration slope, intercept and Brier score. The Youden index identified the optimal probability cut-off. A multivariable logistic-regression model provided interpretable adjusted ORs, with model fit summarised by the Nagelkerke R^2 and the Hosmer–Lemeshow goodness-of-fit test. The clinical impact of actionable, dichotomised predictors was expressed as the number needed to treat (NNT = 1/absolute risk reduction). Multicollinearity among correlated ovarian-reserve markers (AMH, AFC, oocyte yield) was assessed by variance-inflation factors, and non-linearity of continuous predictors was examined with restricted cubic splines. Beyond discrimination, clinical utility was evaluated with positive and negative predictive values at the chosen threshold and with decision-curve (net-benefit) analysis across plausible threshold probabilities. Optimism was quantified by comparing apparent (training) with internal-validation performance, supplemented by bootstrap internal validation. Risk of bias was self-appraised with PROBAST. Absolute risk differences are reported alongside odds ratios because the common outcome (54.6%) makes odds ratios overstate relative risks; the number-needed quantities derived from dichotomised, non-randomised predictors are presented as illustrative, association-based estimates rather than causal treatment effects. Interpretability of the best model was assessed with SHAP summary, importance and dependence plots[17]. Pre-specified subgroup analyses examined maternal age, BMI category, AMH tertile and COS protocol; metric selection followed reproductive-medicine recommendations[15].

Data quality and missing data

Data completeness was audited before analysis; records missing more than 10% of variables were excluded a priori, and remaining sporadic gaps (predominantly in endocrine assays) were addressed by k-nearest-neighbours imputation fitted within the training fold only, to prevent information leakage into the validation set. Continuous predictors were standardised using training-set parameters. Categorical predictors were one-hot encoded. To guard against optimism, all preprocessing, imputation and hyperparameter tuning were nested within the cross-validation performed on the training partition, and the 30% internal-validation partition was accessed only once for final performance estimation.

3. Results and Discussion

Results

Of 1,245 women with PCOS meeting eligibility criteria, the overall cumulative live birth rate was 54.6% (680/1,245; Wilson 95% CI 51.8–57.4%).

Cumulative live birth accrued progressively across sequential transfers—36.0% after the fresh transfer, rising to 47.2%, 52.8% and 54.6% after the first, second and third frozen–thawed transfers respectively. The cohort was predominantly the classic hyperandrogenic-anovulatory phenotype (Rotterdam phenotype A, 58.2%; phenotype B, 11.4%; phenotype C, 18.9%; phenotype D, 11.5%); the multiple-pregnancy rate among live births was 12.4%. Baseline characteristics stratified by outcome are summarised in Table 1.

Women who achieved live birth were younger (30.2 ± 3.5 versus 32.8 ± 4.0 years; Cohen's d = 0.69; p<0.001) and had lower BMI (25.8 ± 3.9 versus 27.3 ± 4.4 kg/m²; d = 0.36; p<0.001). They also had higher AMH (8.2 ± 3.2 versus 7.3 ± 3.8 ng/mL; p=0.041), higher antral follicle counts (19.2 ± 5.8 versus 17.6 ± 6.5; p=0.012), more oocytes retrieved (15.6 ± 5.1 versus 12.5 ± 5.6; d = 0.58; p<0.001) and, most strikingly, more top-quality embryos (4.2 ± 1.6 versus 2.4 ± 1.5; d = 1.16; p<0.001). The antagonist protocol was used in 76.0% of cycles without a significant association with outcome (OR 1.27, 95% CI 0.97–1.65; p=0.075).

Table 1. Baseline maternal characteristics by cumulative live birth (CLBR), with effect sizes.

Variable	Total (n=1245)	CLBR yes (n=680)	CLBR no (n=565)	Effect size	p
Age (years)	31.4 ± 3.8	30.2 ± 3.5	32.8 ± 4.0	d = 0.69	<0.001
BMI (kg/m²)	26.5 ± 4.2	25.8 ± 3.9	27.3 ± 4.4	d = 0.36	<0.001
AMH (ng/mL)	7.8 ± 3.5	8.2 ± 3.2	7.3 ± 3.8	d = 0.26	0.041
Antral follicle count	18.5 ± 6.2	19.2 ± 5.8	17.6 ± 6.5	d = 0.26	0.012
Antagonist protocol, n (%)	946 (76.0)	530 (77.9)	416 (73.6)	OR 1.27 (0.97–1.65)	0.075
Oocytes retrieved (n)	14.2 ± 5.5	15.6 ± 5.1	12.5 ± 5.6	d = 0.58	<0.001
Top-quality embryos (n)	3.4 ± 1.8	4.2 ± 1.6	2.4 ± 1.5	d = 1.16	<0.001

Notes: Continuous variables mean ± SD; categorical n (%); d = Cohen's d.

On dichotomised bivariate analysis, as detailed in Table 2, having ≥3 top-quality embryos was the strongest single correlate of live birth (crude OR 3.18, 95% CI 2.49–4.06; p<0.001), followed by maternal age <35 years (OR 2.11, 95% CI 1.62–2.75; p<0.001) and retrieval of ≥10 oocytes (OR 2.04, 95% CI 1.58–2.63; p<0.001). BMI below 25 kg/m² (OR 1.62, 95% CI 1.29–2.04; p<0.001), AFC ≥18 (OR 1.40; p=0.003) and above-median AMH (OR 1.34; p=0.011) were also significant.

Table 2. Bivariate associations of dichotomised predictors with cumulative live birth.

Predictor (dichotomised)	Crude OR (95% CI)	p
Top-quality embryos ≥3 (vs <3)	3.18 (2.49–4.06)	<0.001
Oocytes retrieved ≥10 (vs <10)	2.04 (1.58–2.63)	<0.001
Maternal age <35 y (vs ≥35)	2.11 (1.62–2.75)	<0.001
BMI <25 kg/m² (vs ≥25)	1.62 (1.29–2.04)	<0.001
AFC ≥18 (vs <18)	1.40 (1.12–1.75)	0.003
AMH above median (vs below)	1.34 (1.07–1.68)	0.011

All five ML algorithms discriminated above chance, but performance increased monotonically from logistic regression to the tree ensembles (Table 3; Figure 1 and Figure 2). XGBoost achieved the best discrimination, with an AUC of 0.82 (DeLong 95% CI 0.79–0.85), accuracy 76.4%, sensitivity 78.2%, specificity 74.1% and an F1-score of 0.78. Gradient boosting (AUC 0.80) and random forest (AUC 0.79) followed, while logistic regression (AUC 0.72, 95% CI 0.69–0.75) and the support vector machine (AUC 0.74) trailed. The improvement of XGBoost over logistic regression was statistically significant (ΔAUC 0.10; DeLong p<0.001). XGBoost was well calibrated (calibration slope 0.97, intercept −0.03, Brier score 0.176), and the Youden-optimal cut-off yielded sensitivity 78.2% and specificity 74.1% (Youden J = 0.523).

Table 3. Discrimination of the five machine-learning models (internal-validation set).

Model	AUC-ROC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1
Logistic regression	0.72 (0.69–0.75)	68.5	70.2	66.5	0.69
Support vector machine	0.74 (0.71–0.77)	70.2	71.5	68.6	0.71
Random forest	0.79 (0.76–0.82)	74.8	76.0	73.3	0.75
Gradient boosting	0.80 (0.77–0.83)	75.5	77.1	73.5	0.76
XGBoost	0.82 (0.79–0.85)	76.4	78.2	74.1	0.78

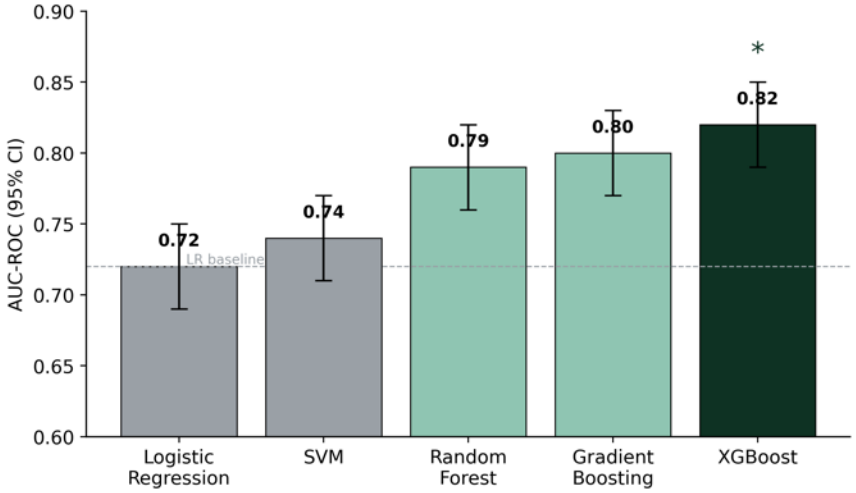

*Figure 1. Discrimination (AUC, 95% CI) of the five models; *p<0.001 for XGBoost versus logistic regression.*

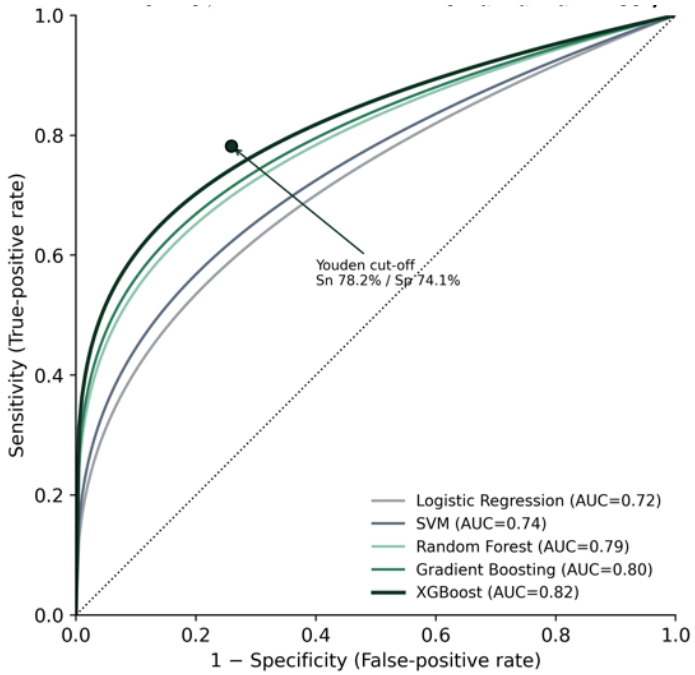

Figure 2. ROC curves on the internal-validation set with the XGBoost Youden cut-off.

Discrimination gains were accompanied by stable calibration across the predicted-risk range: observed live-birth frequencies tracked predicted probabilities closely across risk deciles, with no systematic over- or under-estimation (calibration slope 0.97). Net reclassification favoured XGBoost over logistic

regression, particularly in the intermediate predicted-risk band (predicted probability 0.40–0.60), where the largest counselling uncertainty resides and where capturing non-linear interactions is most valuable. Performance was consistent across both contributing centers, supporting the robustness of the model to inter-center variation in laboratory and stimulation practice. At the Youden-optimal threshold the positive predictive value was 78.9% and the negative predictive value 73.2% at the cohort prevalence of 54.6%, and decision-curve analysis showed a higher net benefit for XGBoost than for logistic regression or a treat-all strategy across the clinically relevant threshold-probability range of 0.35–0.70. Apparent and internal-validation AUCs differed by only 0.02 (0.84 versus 0.82), and bootstrap internal validation gave an optimism-corrected AUC of 0.81, indicating limited overfitting. Variance-inflation factors for the reserve markers were modest (all <3.1), and restricted-cubic-spline modelling confirmed the non-linear, plateauing effect of oocyte yield. When neonatal outcomes were restricted to singletons, mean birth weight was 3145 ± 470 g and the preterm-birth rate was 9.8%, with NICU admission of 7.6%.

Considering the five models in full (Table 3), accuracy rose in step with discrimination: logistic regression 68.5%, support vector machine 70.2%, random forest 74.8%, gradient boosting 75.5% and XGBoost 76.4%, with corresponding F1-scores of 0.69, 0.71, 0.75, 0.76 and 0.78. Sensitivity and specificity were balanced for the best model (78.2% and 74.1%), an important property for a counselling tool that should neither systematically overstate nor understate the chance of success. Across pre-specified subgroups the direction and significance of the leading predictors were preserved, indicating that the model's behaviour was not driven by a single stratum of the cohort.

SHAP analysis rendered the XGBoost model interpretable (Figure 3). The number of top-quality embryos was the dominant positive contributor to predicted CLBR, followed by AMH and oocyte yield; maternal age and BMI exerted the strongest negative effects. Consistent with the bivariate plateau observed clinically, the marginal SHAP contribution of additional oocytes diminished beyond approximately 18 oocytes, reflecting the balance between ovarian response and oocyte quality in PCOS. Translating these findings into clinical terms, women with ≥3 versus <3 top-quality embryos had a CLBR of 68.2% versus 38.1% (absolute risk reduction 30.1%; NNT 3.3).

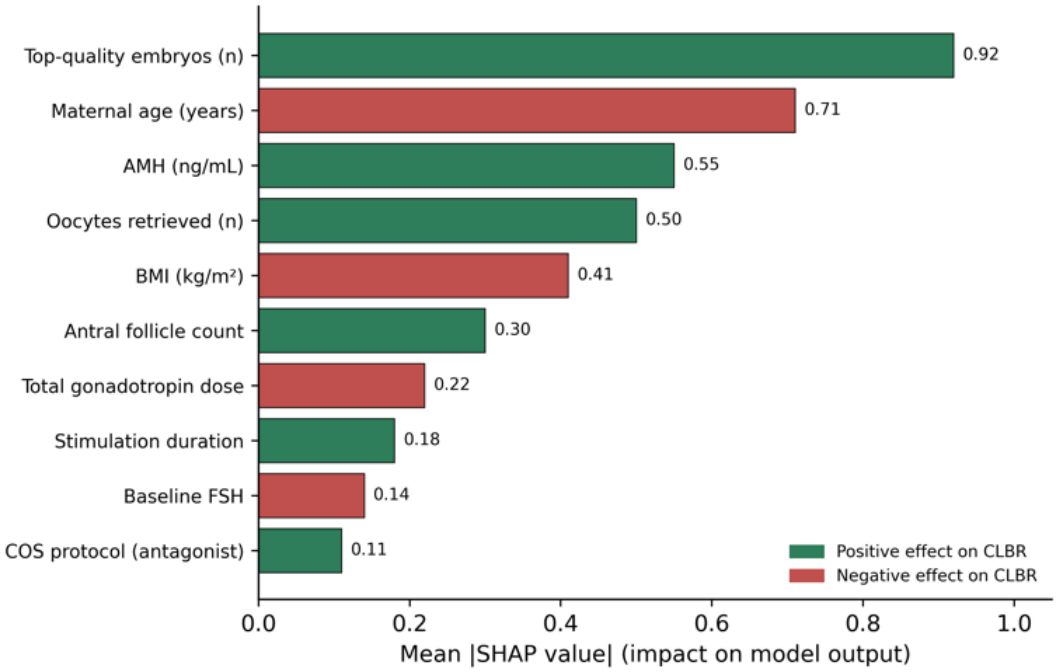

Figure 3. SHAP global feature importance for the XGBoost model.

In the multivariable logistic-regression benchmark (detailed in Table 4 and visualised in Figure 4), independent predictors of cumulative live birth were the number of top-quality embryos (adjusted OR 1.85 per embryo, 95% CI 1.66–2.06; p<0.001), oocytes retrieved (aOR 1.07 per oocyte, 95% CI 1.04–1.10; p<0.001) and AMH (aOR 1.06 per ng/mL, 95% CI 1.01–1.11; p=0.018), whereas BMI (aOR 0.94 per kg/m², 95% CI 0.91–0.97; p<0.001) and maternal age (aOR 0.88 per year, 95% CI 0.85–0.92; p<0.001) were inversely associated. After adjustment, AFC and COS protocol were no longer significant. The model showed acceptable fit (Nagelkerke R^2 = 0.34; Hosmer–Lemeshow x^2 = 7.42, p=0.492).

Table 4. Multivariable logistic-regression adjusted odds ratios for cumulative live birth (Nagelkerke R^2=0.34; Hosmer–Lemeshow p=0.492).

Predictor	Adjusted OR (95% CI)	p
Top-quality embryos (per embryo)	1.85 (1.66–2.06)	<0.001
Oocytes retrieved (per oocyte)	1.07 (1.04–1.10)	<0.001
AMH (per ng/mL)	1.06 (1.01–1.11)	0.018
Antagonist protocol (vs agonist)	1.18 (0.89–1.56)	0.252
Antral follicle count (per unit)	1.02 (0.99–1.05)	0.142
BMI (per kg/m²)	0.94 (0.91–0.97)	<0.001
Maternal age (per year)	0.88 (0.85–0.92)	<0.001

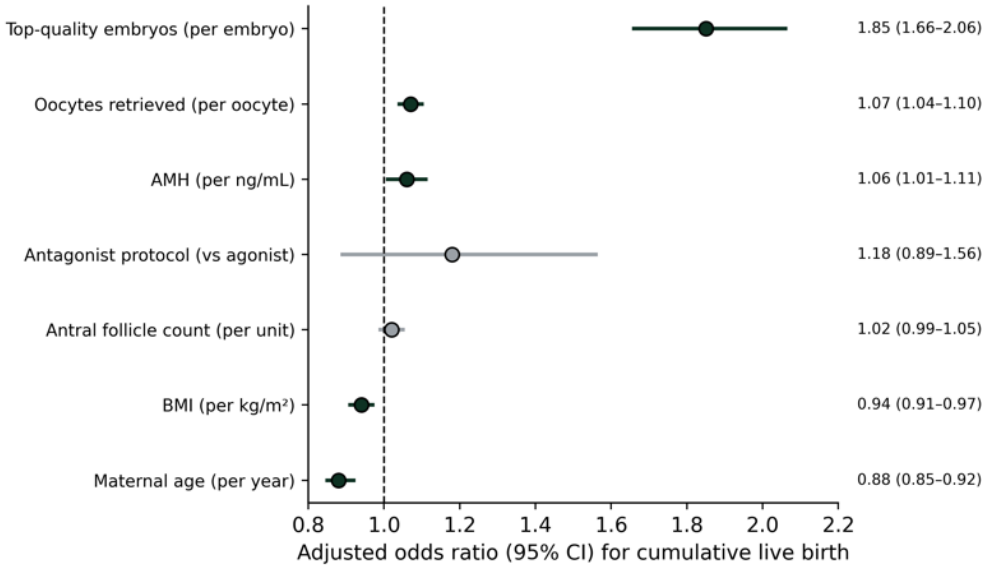

Figure 4. Forest plot of multivariable adjusted odds ratios for cumulative live birth.

Among the 680 women who achieved live birth, perinatal outcomes differed by transfer strategy (Table 5 and Figure 5). A freeze-all approach markedly reduced moderate/severe OHSS (1.2% versus 8.1% with fresh transfer; NNT 14.5 to prevent one case), but was associated with more hypertensive disorders of pregnancy (15.7% versus 10.9%; OR 1.52, 95% CI 1.01–2.29; p=0.044) and higher mean birth weight (3120 ± 505 versus 2985 ± 540 g; Δ +135 g; p=0.002). Caesarean delivery (48.5% overall), gestational diabetes, preterm birth, low birth weight, 5-minute Apgar <7 and NICU admission did not differ significantly between strategies. Pre-specified subgroup CLBRs were 59.8% in women aged <35 years versus 41.3% at ≥35 years, 60.9% at BMI <25 versus 48.7% at ≥25 kg/m², and 58.4% versus 50.1% across upper versus lower AMH tertiles.

Table 5. Maternal and neonatal outcomes among live births, by transfer strategy (% unless stated).

Outcome	Fresh ET	Freeze-all/FET	Effect / NNT	p
Moderate/severe OHSS	8.1	1.2	NNT 14.5	<0.001
Hypertensive disorders of pregnancy	10.9	15.7	OR 1.52 (1.01–2.29)	0.044
Gestational diabetes	16.8	19.4	OR 1.19 (0.81–1.75)	0.372
Caesarean delivery	45.2	51.3	OR 1.28 (0.95–1.72)	0.108
Preterm birth <37 weeks	16.1	12.0	OR 0.71 (0.46–1.10)	0.124
Birth weight (g)	2985 ± 540	3120 ± 505	Δ +135 g	0.002
Low birth weight <2500 g	13.4	8.9	OR 0.63 (0.39–1.02)	0.060
NICU admission	11.2	8.1	OR 0.70 (0.43–1.14)	0.151

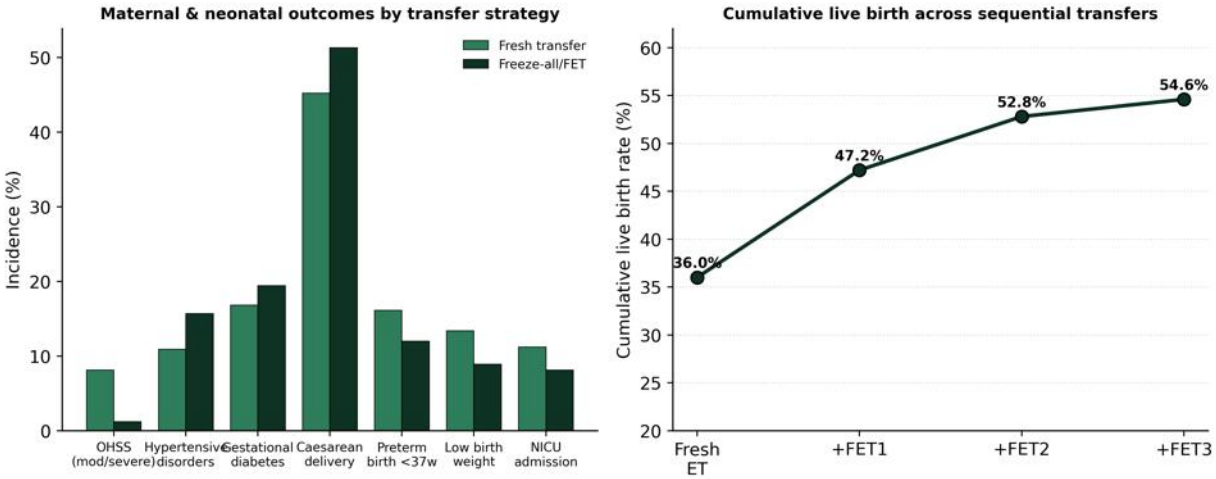

Figure 5. Maternal and neonatal outcomes by transfer strategy and cumulative live birth across sequential transfers.

Discussion

In this multicenter cohort of 1,245 Indonesian women with PCOS undergoing a first IVF/ICSI cycle, an explainable XGBoost model predicted cumulative live birth with good discrimination (AUC 0.82) and was well calibrated, significantly outperforming conventional logistic regression (AUC 0.72; ΔAUC 0.10, p<0.001). SHAP analysis identified the number of top-quality embryos, maternal age, AMH and oocyte yield as the dominant predictors, and translated the model into clinically intuitive terms—each additional top-quality embryo raised the adjusted odds of live birth by 85%, and the threshold of three top-quality embryos corresponded to a number-needed-to-treat of just 3.3 for an additional live birth.

Our discrimination is concordant with, and at the upper end of, the published range for IVF outcome prediction. Pre-treatment ML models for live birth have reported AUCs of roughly 0.70–0.80 in general IVF populations, and internal comparisons consistently show tree-ensemble methods edging out regression by capturing non-linear interactions.[12,15,16] The incremental 0.10 AUC gain of XGBoost over logistic regression we observed mirrors these reports and is clinically meaningful at the decision threshold, where it corresponds to materially better separation of likely responders from non-responders.[12,15]

The pre-eminence of embryo quality in both our SHAP rankings and multivariable model aligns with a broad literature establishing embryo morphology and developmental competence as the proximal determinants of implantation and live birth, and with the rapid maturation of AI-based embryo assessment[6,19,20]. That oocyte yield mattered but with

a diminishing—and ultimately plateauing—marginal effect beyond approximately 18 oocytes is biologically coherent for PCOS, in which exuberant follicular recruitment is offset by compromised oocyte competence and OHSS risk; it also cautions against equating high oocyte numbers with proportionally higher success[5,6].

The inverse contributions of maternal age and BMI are expected but clinically actionable. Age remains the strongest non-modifiable determinant of ART success, whereas BMI is modifiable; our adjusted estimate (aOR 0.94 per kg/m²) supports pre-treatment weight optimisation as a lever to improve outcomes in overweight and obese women with PCOS[7,8]. AMH retained independent predictive value, consistent with its role as a quantitative marker of the PCOS ovarian-reserve phenotype, although its predictive ceiling reflects that quantity does not guarantee quality[5,10].

A distinctive strength of our approach is interpretability. The principal barrier to clinical adoption of AI in reproductive medicine has been the 'black-box' problem, repeatedly highlighted in critical appraisals of fertility AI[13,17,18]. By coupling a high-performing ensemble with SHAP, we provide patient-level, directionally explicit explanations that clinicians can interrogate during counselling—an approach already demonstrated for other reproductive-medicine prediction tasks such as ovarian-response assessment.[17,21] This transparency, combined with reporting aligned to TRIPOD+AI, addresses two of the most frequent methodological criticisms of prior IVF prediction models.[12,15]

Clinically, the model could be embedded within hospital information systems via an application-programming interface to generate individualised CLBR estimates at the point of care. Such estimates would support realistic counselling, shared decisions about the number of cycles to pursue, and selective application of strategies such as freeze-all or weight reduction before transfer[7,14]. Our perinatal sub-analysis reinforces that these decisions carry trade-offs: a freeze-all strategy substantially reduced OHSS (NNT 14.5) but was associated with more hypertensive disorders and heavier infants, echoing randomised and cohort evidence and underscoring that live birth, while paramount, must be weighed against maternal and neonatal safety[22,23].

From an Indonesian and broader Southeast Asian perspective, our findings help close a recognised evidence gap, since most ART prediction tools derive from European and North American cohorts and may transfer poorly to regional populations and practice patterns[2,9]. A locally derived, PCOS-specific tool is therefore of particular relevance to the region's rapidly expanding ART sector, where demand frequently outstrips capacity and accurate prognostication can improve both counselling and resource allocation[10].

This study has several strengths: a multicenter design that reduces single-institution bias, a comprehensive and clinically granular feature set, a sample size that comfortably satisfies events-per-variable requirements, explicit calibration and effect-size reporting, and SHAP-based interpretability that avoids the black-box limitation. The addition of maternal and neonatal outcomes situates a predictive-analytics study within a patient-safety frame appropriate to obstetric practice.

Limitations must be acknowledged. The retrospective design is susceptible to selection and information bias, and some predictors (for example, fine-grained embryo morphokinetics) were unavailable. The model underwent internal validation only; although calibration was good, performance may attenuate in external populations, and prospective, multicenter external validation with impact assessment is required before deployment[13,18]. The perinatal comparisons were observational and potentially confounded by indication for freeze-all. Finally, as with all prediction tools, the model informs but does not replace individualised clinical judgement.

Our results contribute to an ongoing debate about whether ML genuinely outperforms regression for clinical prediction. Some methodological reviews argue that, for tabular clinical data with modest dimensionality, well-specified regression often matches flexible learners. Our experience is

nuanced: in PCOS—where interactions among embryo number, oocyte yield and age are demonstrably non-additive—the ensemble captured signal that the linear model missed, yet the regression remained a transparent, well-calibrated benchmark whose adjusted odds ratios aid mechanistic interpretation. The two approaches are therefore complementary rather than competing, and reporting both, as we have done, is good practice[12,15].

Several implementation considerations bear emphasis. Deployment of any prediction model demands ongoing surveillance for calibration drift as laboratory techniques, stimulation protocols and embryo-culture systems evolve; a model trained on 2018–2023 data will require periodic recalibration. Equity is also pertinent: models must be monitored to ensure they do not encode or amplify disparities in access to additional cycles. Transparent, SHAP-based explanations and adherence to TRIPOD+AI reporting facilitate the external scrutiny, regulatory review and clinician acceptance that responsible adoption requires[17,18].

From a health-systems perspective, accurate CLBR prediction has tangible value in resource-constrained settings typical of much of Southeast Asia, where ART is largely self-funded and each cycle represents a substantial household expenditure. Reliable individualised estimates can inform shared decisions about whether and how often to proceed, reduce the emotional and financial costs of futile cycles, and support equitable prioritisation where capacity is limited. A locally derived tool is more likely to be trusted and adopted by regional clinicians than an imported model of uncertain transportability[2,9].

It is worth situating the perinatal findings within obstetric priorities. Although the index outcome of this study was live birth, the downstream maternal and neonatal consequences of cycle-management choices are central to obstetric care. The substantial reduction in OHSS with a freeze-all strategy (NNT 14.5) is a clear safety benefit in the OHSS-prone PCOS population, but the accompanying signal toward more hypertensive disorders of pregnancy and heavier infants—consistent with the wider frozen-transfer literature—means that prediction should be embedded within a holistic, safety-aware counselling framework rather than optimising live birth in isolation[22,23]. Integrating predicted CLBR with individualised OHSS and obstetric-risk estimates is a logical next step toward truly personalised ART.

Future work should pursue prospective, multicenter external validation across diverse Indonesian and regional centers, ideally followed by a prospective impact study testing whether model-guided counselling changes patient decisions, cycle uptake and satisfaction. Incorporating richer predictors—time-lapse embryo morphokinetics, endometrial receptivity markers and standardised metabolic phenotyping of PCOS—may further improve discrimination, while federated-learning approaches could enable multi-institutional model training without sharing identifiable patient data, addressing both privacy and generalisability simultaneously[18,19].

Finally, the methodological template demonstrated here—multicenter derivation, an interpretable benchmark, explicit calibration, SHAP-based explanation, effect-size and NNT reporting, and TRIPOD+AI adherence—may serve as a model for prediction studies addressing other reproductive-medicine outcomes, such as OHSS risk, optimal embryo number for transfer, or obstetric complications after ART. Standardising this template across the field would improve comparability, reduce research waste and accelerate the translation of trustworthy algorithms into routine fertility care[13,15].

The patient perspective further motivates this work. For couples facing PCOS-associated infertility, uncertainty is itself a source of distress, and vague prognostic statements can erode trust and complicate consent. A calibrated, individualised probability—accompanied by a transparent account of which of the patient's own characteristics drive that estimate—supports genuinely informed, shared decision-making. The SHAP framework allows a clinician to explain, for example, that a favourable embryo cohort is pulling a particular patient's predicted success upward while elevated BMI is pulling it down, thereby reframing the consultation

around modifiable factors and realistic expectations rather than a single opaque percentage.

Methodologically, we deliberately reported the multivariable logistic-regression model alongside the machine-learning ensemble so that the adjusted odds ratios remain available for mechanistic and biological interpretation, while the ensemble supplies the superior individual-level discrimination. This dual reporting responds directly to long-standing concerns that prediction studies privilege a single headline metric; by presenting discrimination, calibration, effect sizes, number-needed-to-treat and explainability together, the analysis offers a more complete and clinically usable evidence package than discrimination alone would provide[12,15].

Several boundary conditions should temper interpretation. The cohort comprised women aged 20–40 years undergoing a first cycle, so extrapolation to older women, to repeat-cycle populations, or to non-PCOS aetiologies is not supported by these data. Laboratory assays and embryo-grading conventions, although accredited, differ between centers and over time, which is precisely why prospective external validation and periodic recalibration are emphasised. Within these limits, the consistency of the leading predictors with established reproductive biology and with the wider literature lends confidence that the model is capturing genuine prognostic signal rather than center-specific artifact[5,6].

In summary, the present study advances reproductive-medicine prediction in three concrete ways. It delivers a PCOS-specific cumulative-live-birth model rather than a generic IVF model; it pairs a high-performing ensemble with patient-level explainability to overcome the black-box barrier that has impeded clinical uptake; and it draws on multicenter Southeast Asian data that are markedly under-represented in the existing literature, improving the regional relevance of the resulting tool. Taken together with rigorous calibration, effect-size and number-needed-to-treat reporting, and adherence to contemporary reporting standards, these features position the model as a credible candidate for prospective evaluation and, ultimately, for integration into individualised fertility counselling and precision reproductive care[10,13].

4. Conclusion

An explainable XGBoost model, benchmarked against logistic regression and reported to contemporary standards, accurately and interpretably predicts cumulative live birth in women with PCOS undergoing IVF/ICSI on internal validation (AUC 0.82; optimism-corrected 0.81). The number of top-quality embryos, maternal age, AMH and oocyte yield were the dominant predictors, and the model translated into actionable thresholds (NNT 3.3 for ≥3 top-quality embryos). Integrated into clinical workflows, such a tool could enhance individualised counselling and precision care in reproductive endocrinology. Prospective, multicenter external validation and prospective impact studies are warranted before routine clinical adoption.

Declarations

Acknowledgements

None.

Author Contributions

All authors contributed to conceptualisation, methodology, formal analysis, investigation, data curation, and writing (original draft and review/editing), and approved the final manuscript (CRediT).

Conflict of Interest

The authors declare no conflict of interest.

Funding

This research received no specific grant from any funding agency.

Ethics and Consent

The study was approved by CMHC Research Center (approval No. CMHC/EC/2023/0214), in accordance with the Declaration of Helsinki; the requirement for individual informed consent was waived owing to the retrospective, fully de-identified design. No identifiable patient material is included.

5. References

1. Stener-Victorin E, Teede H, Norman RJ, et al. Polycystic ovary syndrome. Nat Rev Dis Primers. 2024;10(1):27. doi:10.1038/s41572-024-00511-3.
2. Louis CM, Handayani N, Aprilliana T, et al. Genetic algorithm-assisted machine learning for clinical pregnancy prediction in in vitro fertilization. AJOG Glob Rep. 2022;3(1):100133. doi:10.1016/j.xagr.2022.100133.
3. Jiang H, Si M, Tian T, et al. Adiposity and lipid metabolism indicators mediate the adverse effect of glucose metabolism indicators on oogenesis and embryogenesis in PCOS women undergoing IVF/ICSI cycles. Eur J Med Res. 2023;28(1):216. doi:10.1186/s40001-023-01174-8.
4. He Y, Li R, Yin J, et al. Influencing of serum inflammatory factors on IVF/ICSI outcomes among PCOS patients with different BMI. Front Endocrinol (Lausanne). 2023;14:1204623. doi:10.3389/fendo.2023.1204623.
5. Su N, Zhan J, Xie M, et al. High anti-Mullerian hormone level is adversely associated with cumulative live birth rates of two embryo transfers after the first initiated cycle in patients with polycystic ovary syndrome. Front Endocrinol (Lausanne). 2023;14:1123125. doi:10.3389/fendo.2023.1123125.
6. Cubillos-Garcia SP, Revilla-Pacheco F, Meneses-Mayo M, et al. Required number of blastocysts transferred, and oocytes retrieved to optimize live and cumulative live birth rates in the first complete cycle of IVF for autologous and donated oocytes. Arch Gynecol Obstet. 2024;310(5):2681-2690. doi:10.1007/s00404-024-07712-x.
7. Fouks Y, Neuhausser W, Ryley D, et al. ART outcomes in lean compared to obese phenotypes of polycystic ovarian syndrome. J Assist Reprod Genet. 2023;40(6):1437-1445. doi:10.1007/s10815-023-02804-0.
8. Li J, Shi H, Bu Z, et al. Effect of body mass index on the cumulative live birth rate over multiple complete IVF cycles in women with polycystic ovary syndrome: a retrospective study. Obes Res Clin Pract. 2023;17(2):130-136. doi:10.1016/j.orcp.2023.03.001.
9. Kobayashi M, Asada Y, Sugishita Y, et al. Cumulative ongoing pregnancy and live birth rates after assisted reproduction among women with diminished ovarian reserve: a multi-institutional retrospective cohort study. J Obstet Gynaecol Res. 2025;51(6):e16336. doi:10.1111/jog.16336.
10. Nguyen DK, O'Leary S, Gadalla MA, et al. The predictive value of anti-Mullerian hormone for natural conception leading to live birth in subfertile couples. Reprod Biomed Online. 2022;44(3):557-564. doi:10.1016/j.rbmo.2021.11.018.
11. Liu S, Zhou X, Jie H, et al. Higher cumulative live birth rate but also higher late miscarriage risk in non-obese women with polycystic ovary syndrome undergoing the first IVF/ICSI cycle. Int J Womens Health. 2024;16:289-298. doi:10.2147/IJWH.S445021.
12. Liu X, Chen Z, Ji Y, et al. Construction of the machine learning-based live birth prediction models for the first in vitro fertilization pregnant women. BMC Pregnancy Childbirth. 2023;23(1):476. doi:10.1186/s12884-023-05775-3.
13. Zhu S, Huang Z, Chen X, et al. Construction and evaluation of machine learning-based prediction model for live birth following fresh embryo transfer in IVF/ICSI patients with polycystic ovary syndrome. J Ovarian Res. 2025;18(1):70. doi:10.1186/s13048-025-01654-x.
14. Deepika K, Suvarna R, Sumi M, et al. HCG trigger versus GnRH agonist trigger in PCOS patients undergoing IVF cycles: frozen embryo transfer outcomes. JBRA Assist Reprod. 2021;25(1):48-58. doi:10.5935/1518-0557.20200028.
15. Peng J, Geng X, Zhao Y, et al. Machine learning algorithms in constructing prediction models for assisted reproductive technology (ART) related live birth outcomes. Sci Rep. 2024;14(1):32083. doi:10.1038/s41598-024-83781-x.
16. Zhang Y, Shen L, Yin X, et al. Live-birth prediction of natural-cycle in vitro fertilization using 57,558 linked cycle records: a machine learning perspective. Front Endocrinol (Lausanne). 2022;13:838087.

doi:10.3389/fendo.2022.838087.

17. Hsu WH, Ko AT, Weng CS, et al. Explainable machine learning model for predicting skeletal muscle loss during surgery and adjuvant chemotherapy in ovarian cancer. J Cachexia Sarcopenia Muscle. 2023;14(5):2044-2053. doi:10.1002/jcsm.13282.
18. Wan X, Yu M, Wu X, et al. Machine learning prediction of clinical pregnancy in endometriosis patients following fresh IVF/ICSI-ET. Eur J Med Res. 2025;30(1):838. doi:10.1186/s40001-025-03113-1.
19. Liu H, Chen L, Shan G, et al. An interpretable artificial intelligence approach to differentiate between blastocysts with similar or same morphological grades. Hum Reprod. 2025;40(6):1077-1086. doi:10.1093/humrep/deaf066.
20. Yang L, Peavey M, Kaskar K, et al. Development of a dynamic machine learning algorithm to predict clinical pregnancy and live birth rate with embryo morphokinetics. F S Rep. 2022;3(2):116-123. doi:10.1016/j.xfre.2022.04.004.
21. Wei J, Xiong D, Zhang Y, et al. Predicting ovarian responses to the controlled ovarian hyperstimulation in elderly infertile women using clinical measurements and random forest regression. Eur J Obstet Gynecol Reprod Biol. 2023;288:153-159. doi:10.1016/j.ejogrb.2023.07.012.
22. Pinborg A, Wennerholm UB, Bergh C, et al. Long-term outcomes for children conceived by assisted reproductive technology. Fertil Steril. 2023;120(3):449-456. doi:10.1016/j.fertnstert.2023.04.022.
23. Yu T, Wu D, Cao Y, et al. Association between menstrual patterns and adverse pregnancy outcomes in patients with polycystic ovary syndrome. Front Endocrinol (Lausanne). 2021;12:740377. doi:10.3389/fendo.2021.740377.