scientific reports

Check for updates

OPEN

Development and validation of machine learning models for predicting blastocyst yield in IVF cycles

Wen-jie Huo[1], Fei Peng[2,3], Song Quan[1] & Xiao-cong Wang[1]✉

Predicting blastocyst formation poses significant challenges in reproductive medicine and critically influences clinical decision-making regarding extended embryo culture. While previous research has primarily focused on determining whether an IVF cycle can produce at least one blastocyst, less attention has been given to quantifying blastocyst yields. This study aims to develop and validate such a quantitative predictive tool for IVF cycles. We employed three machine learning models—SVM, LightGBM, and XGBoost—which demonstrated comparable performance and outperformed traditional linear regression models (R^2: 0.673–0.676 vs. 0.587, Mean absolute error: 0.793–0.809 vs. 0.943). Ultimately, LightGBM emerged as the optimal model, due to utilizing fewer features (8 vs. 10–11 in SVM/XGBoost) and offering superior interpretability. We then stratified predictions and actual yields into three categories (0, 1–2, and ≥ 3 blastocysts) to evaluate the model's discriminative performance. In this multi-classification task, LightGBM demonstrated robust accuracy (0.675–0.71) with fair-to-moderate agreement (kappa coefficients: 0.365–0.5) across both the overall cohort and poor-prognosis subgroups. Feature importance analysis identified three critical predictors: the number of extended culture embryos, the mean cell number on Day 3, and the proportion of 8-cell embryos. By leveraging the potential of machine learning, this research provides clinicians with valuable insights for making individualized decisions regarding extended embryo culture.

Keywords Blastocyst yield, Clinical decision support, Extended embryo culture, In vitro fertilization, Machine learning

In vitro fertilization-embryo transfer (IVF-ET) has transformed the landscape of infertility treatment, offering hope to millions of couples worldwide. Advances in culture conditions have increasingly supported extending embryo culture to Day 5 or 6 (blastocyst stage), rather than the conventional Day 2 or 3 (cleavage stage), in an effort to enhance the live birth rates[1]. However, this strategy also raises concerns about the reduction in available embryos[2] and its further impact on the cumulative success rate of each IVF cycle[3,4] due to suboptimal culture conditions in vitro. Consequently, despite the potential benefits of extended culture, its overall safety and efficacy warrant careful scrutiny[5].

To enhance the safety, some broad recommendations have been offered: American Society for Reproductive Medicine (ASRM) supports blastocyst culture for good-prognosis patients[2]and National Collaborating Centre for and Children's (NICE) advocates cleavage-stage transfer for patients with limited embryos[6]. However, these recommendations focus on singular or narrowly defined factors. There is a wide variation (0-100%) in blastocyst formation rates among patients[7]which heightens the risk of losing potentially viable embryos. In practice, multiple patient-specific prognostic elements must be evaluated simultaneously. There is a clear clinical need for a cycle-based model to predict blastocyst yield, which would support individualized risk-benefit assessments and enhance the understanding of patient heterogeneity, ultimately improving the precision of extended embryo culture strategies.

Existing cycle-based models have been limited to binary prediction of complete culture failure[8–10]. These tools overlook the availability of surplus embryos after transferring a viable blastocyst, an essential consideration when determining whether to pursue extended embryo culture[1].Although few studies have identified several potential determinants of blastocyst yield (e.g., oocyte number, day-3 embryo quality) using traditional statistical

[1]Department of Obstetrics and Gynecology, Nanfang Hospital, Southern Medical University, Guangzhou, China. [2]Department of Psychology, School of Public Health, Southern Medical University, Guangzhou, China. [3]Department of Psychiatry, Zhujiang Hospital, Southern Medical University, Guangzhou, China. ✉email: xwang@smu.edu.cn

methods[11,12]complex interactions among these parameters may limit the predictive accuracy of such models. Machine learning, known for its ability to capture nonlinear relationships[13]has not yet been applied to blastocyst yield prediction. To address this gap, we aimed to develop and internally validate machine learning models to quantitatively predict blastocyst yields.

We set out to create models that are clinically acceptable in performance, methodologically transparent, and reliable for supporting decisions on extended embryo culture. Following the TRIPOD + AI guidelines for clinical prediction model reporting[14]we analyzed over 9,000 IVF/intracytoplasmic sperm injection (ICSI) cycles using a structured and transparent methodology. Potential clinical predictors were incorporated to establish the initial feature set, and the dataset was randomly split into training and testing subsets. We trained three machine learning models alongside a baseline linear regression model using backward feature selection, iteratively removing the least informative features from the maximal set. Internal validation was performed on the testing set with multiple performance metrics to assess robustness in our center.

Results

Dataset characteristics

A total of 9,649 cycles were included in our dataset, of which 3,927 (40.7%) produced no usable blastocysts, 3,633 (37.7%) yielded one or two usable blastocysts, and 2,089 cycles (21.6%) resulted in three or more usable blastocysts. This dataset was randomly split into training and test sets (see Methods). Potential features and cycle-based outcome for both sets are summarized in Table 1.

Model evaluation

We performed model-RFE analysis to identify the optimal feature subset and compare the performance of the models (Fig. 1). The RFE results showed that all models maintained stable performance with 8 to 21 features. A sharp decline in R^2 values occurred during the final phase of feature selection, when the number of features reduced to 6 or fewer. The MAE curves followed a similar trend. Table 2 provides a comparative summary of the optimal performance metrics across all models.

The three machine learning models had remarkably similar patterns, achieving optimal performance with approximately 8 to 11 features, reaching R^2 values of 0.67–0.68 and MAE values of 0.79–0.81. In contrast, the linear regression model consistently underperformed, with an R^2 of 0.59 and an MAE of 0.94. Among the three machine learning models with comparable performance, LightGBM was selected as the preferred choice due to its use of fewer features (8 versus 10–11 for SVM and XGBoost), which reduces overfitting risk and enhances simplicity for clinical application. It also offers greater interpretability than the complex kernel transformations used in SVM. Overall, LightGBM provides the best balance of accuracy, practicality and interpretability, making it well-suited for clinical decision support.

Subgroup characteristics by blastocyst yield

Given that poor prognosis patients face more urgent decision-making dilemmas during extended culture due to inherently lower blastocyst yields, we focused our subgroup analysis in this population. Figure 2 illustrates the three-class distribution of predicted versus actual blastocyst yields (0, 1–2, and ≥ 3 blastocysts) in the overall test sets and specific subgroups, with confusion matrices presented as bar plots. In the overall cohort, the predicted yields are noticeably skewed in favor of the actual yields. However, in the subgroups, this trend varies when actual yields are ≥ 3, as predictions tend to fall into the lower categories (0 and 1–2), suggesting that the model may underestimate yields in these subgroups. Fortunately, the proportion of actual yields ≥ 3 is relatively low in these subgroups, at 13.4% for advanced maternal age, 8.3% for poor embryo morphology, and 2.5% for low embryo count, compared to 21.6% in the overall test sets.

Table 3 presents the comprehensive evaluation metrics of our model on three-class classification. For the overall cohort, it achieved an accuracy of 0.678 and a Kappa coefficient of 0.5. When compared to the overall cohort, the three subgroups exhibited small variation in accuracy, ranging from 0.675 to 0.71, while the Kappa coefficients showed a decrease, ranging from 0.365 to 0.472. Additionally, F1(0) scores increased, whereas the F1 scores for the 1–2 and ≥ 3 blastocyst cases declined.

Model interpretation

Feature importance analysis revealed the primary predictors of blastocyst yiled (Fig. 3A). The LightGBM model identified eight key features, with the number of extended culture embryos emerging as the most critical predictor (61.5%). Other predictors included Day 3 embryo-related metrics: mean cell number (10.1%), the proportion of 8-cell embryos (10.0%), the proportion of symmetry (4.4%), and mean fragmentation (2.7%), while Day 2 characteristics—the proportion of 4-cell embryos (7.1%)—also contributed substantially. Demographic and treatment-related factors, including female age (2.4%) and the number of 2PN embryos (1.7%), demonstrated relatively lower importance in predicting blastocyst development.

Individual conditional expectation (ICE) and partial dependence plots (Fig. 3B) elucidated how the top six features modulated model predictions. The number of extended culture embryos, mean cell number (D3), proportion of 8-cell embryos (D3), proportion of symmetry (D3), and proportion of 4-cell embryos (D2) positively influenced blastocyst yield, while fragmentation negatively impacted it. Although these general trends were evident, substantial variability in individual predictions at specific feature values underscores that blastocyst yield results from a complex interplay of multiple factors rather than being determined by a single predictor.

Characteristic	Training set	Test set
Cycles, n	6756	2893
Female age, year	33.5 ± 5.2	33.5 ± 5.2
Male age, year	35.9 ± 6.2	35.9 ± 6.1
BMI, kg/m^2	21.6 ± 2.9	21.8 ± 2.9
Infertility type		
Primary	36.7%	37.6%
Secondary	63.3%	62.4%
Infertility cause		
Female tubal factors	59.9%	59.1%
Other female factors (with or without tubal involvement)	11.7%	13.2%
Male factors	16.1%	14.9%
Combined factors	5.2%	5.2%
Unexplained causes	7.0%	7.5%
Stimulation protocol		
Agonist	43.5%	43.0%
Antagonist	43.4%	44.1%
Minimal stimulation	9.5%	9.6%
Others	3.6%	3.2%
Total gonadotropin dose, IU	2153.0 ± 881.6	2174.2 ± 893.7
Stimulation duration, day	10.1 ± 2.4	10.2 ± 2.5
Number of oocytes retrieved	11.9 ± 6.6	12.0 ± 6.7
Fertilization method		
IVF	74.7%	75.2%
ICSI	13.9%	13.2%
IVF + ICSI	7.0%	7.4%
PESA/TESA	4.4%	4.2%
Number of 2PN embryos	7.4 ± 4.4	7.5 ± 4.4
Number of extended culture embryos	5.2 ± 3.8	5.3 ± 3.8
Embryo proportion, per cycle		
4-cell (D2)	0.38 ± 0.30	0.37 ± 0.30
> 4-cell (D2)	0.29 ± 0.28	0.30 ± 0.28
8-cell (D3)	0.14 ± 0.21	0.14 ± 0.21
> 8-cell (D3)	0.23 ± 0.28	0.24 ± 0.28
< 10%-fragmentation (D3)	0.44 ± 0.37	0.45 ± 0.37
symmetry (D3)	0.77 ± 0.28	0.76 ± 0.28
Mean cell number (D2), per cycle	3.9 ± 0.9	4.0 ± 0.9
Mean cell number (D3), per cycle	6.9 ± 1.6	6.9 ± 1.6
Mean fragmentation (D3), per cycle	10.1 ± 4.7	9.9 ± 4.7
Outcome		
Blastocyst yield, per cycle	1.5 ± 2.0	1.5 ± 2.0
Proportion of cycles without a usable blastocyst	40.5%	41.1%
Proportion of cycle with 1–2 usable blastocyst	37.8%	37.3%
Proportion of cycle with 3 or more usable blastocysts	21.7%	21.6%

Table 1. Patient and cycle characteristics. BMI, body mass index; IVF, in vitro fertilization; ICSI, intracytoplasmic sperm injection; PESA, percutaneous epididymal sperm aspiration; TESA, testicular sperm aspiration; PN, pronuclei; D2, day 2 after fertilization; D3, day 3 after fertilization.

Discussion

In this study, we developed machine learning models to quantitatively predict blastocyst yields in IVF cycles, demonstrating that machine learning algorithms significantly outperformed traditional linear regression. LightGBM emerged as the most effective model, achieving superior predictive performance with fewer features and offering enhanced interpretability compared to other models. We also identified the number of embryos in extended culture and early embryo morphology contributed substantially to the prediction of blastocyst yield. Our research provides a cycle-level perspective for personalized decision-making regarding embryo culture strategies, differing from previous tools that focused on binary predictions for individual embryo 15[18] or per cycle 8[10].

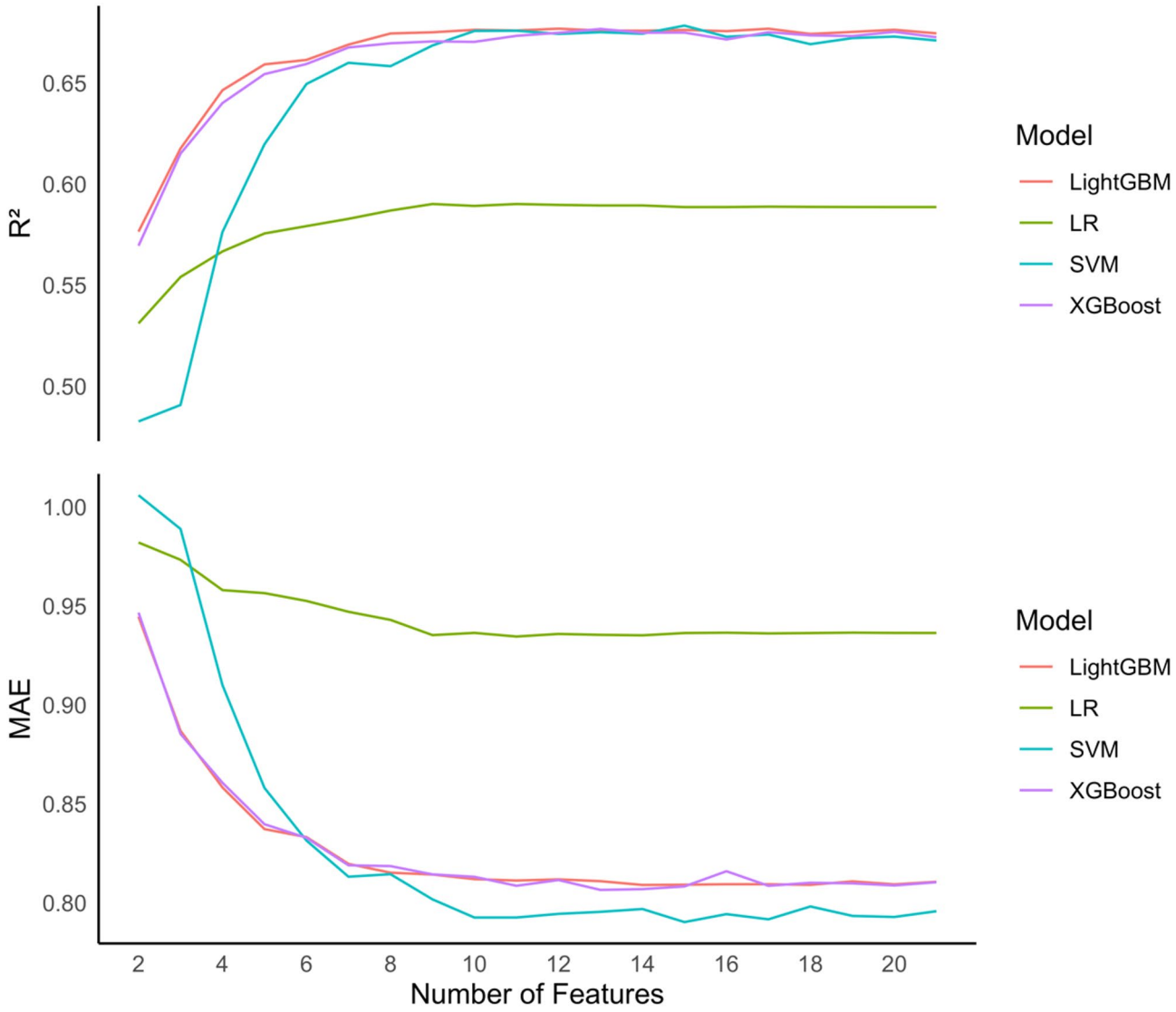

Fig. 1. Performance comparison of machine learning models using recursive feature elimination (RFE). The figure illustrates the impact of RFE on model performance across four machine learning algorithms: Light Gradient Boosting Machine (LightGBM), Linear Regression (LR), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). Features are systematically eliminated from 21 down to 2. The top panel presents the test R^2 (coefficient of determination), where higher values indicate better model fit, while the bottom panel displays the test Mean Absolute Error (MAE), where lower values represent better prediction accuracy.

Models	Feature	R^2	MAE	RMSE
Linear regression	8	0.587	0.943	1.26
SVM	10	0.676	0.793	1.12
Lightgbm	8	0.675	0.813	1.12
XGBoost	11	0.673	0.809	1.12

Table 2. Performance of models. MAE, mean absolute error; RMSE, root mean square error.

The standard evaluation across multiple models and the use of explainable approach here, address critical gaps in prior studies[11,12,19]and align with recent methodological recommendations[14,20,21]. Establishing statistical associations[11,12]does not imply the ability to make generalized predictions; therefore, evidence for prediction requires testing the model on data separate from that used to estimate its parameters[22]. While summing individual embryo probabilities, as suggested by Jiang et al., can also estimate blastocyst yield, this indirect

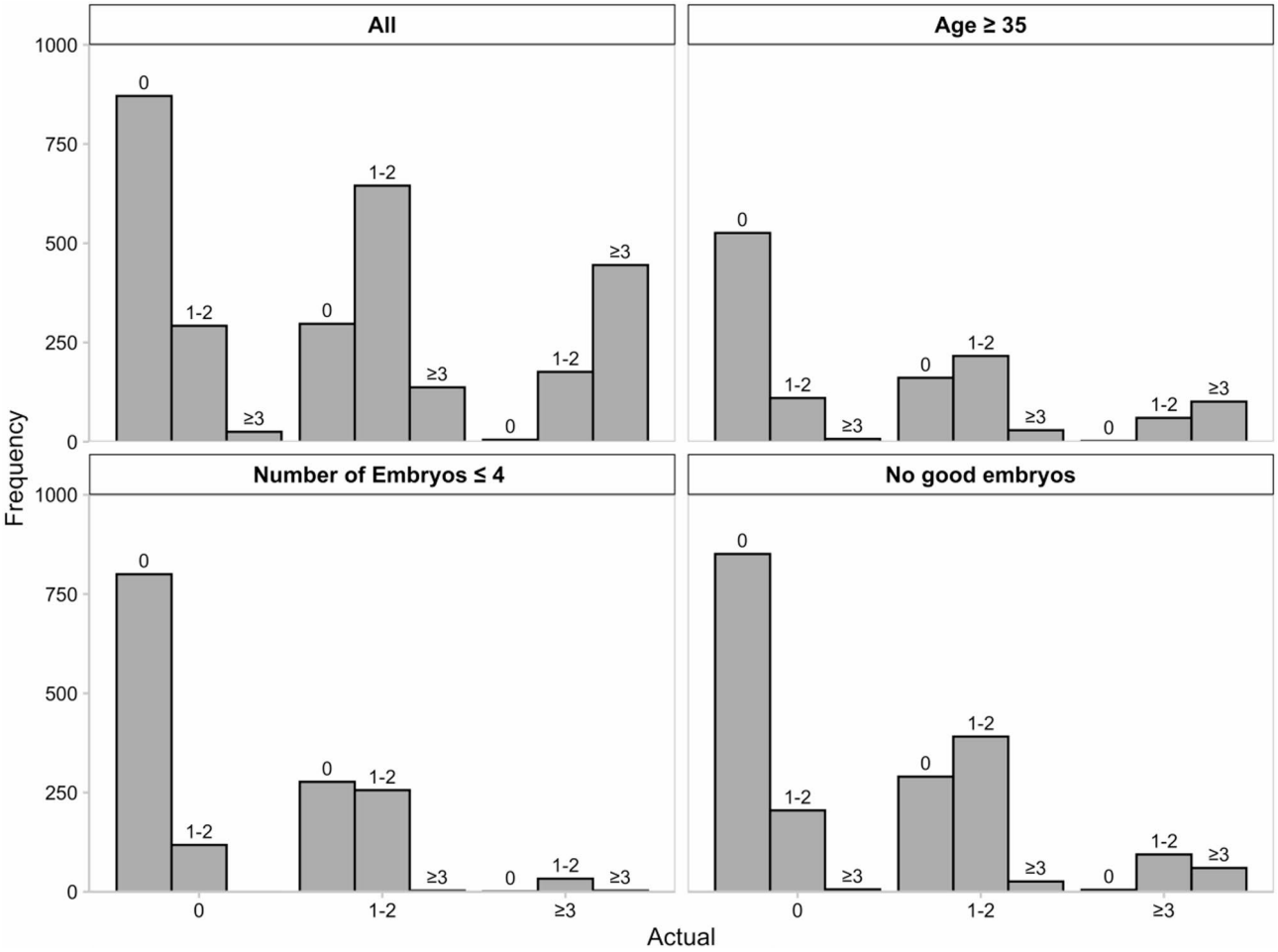

Fig. 2. Distribution of predicted versus actual blastocyst yields in bins across the overall cohort and subgroups. The confusion matrices visualized as bar plots show the relationship between predicted and actual blastocyst yields (0, 1–2, and ≥ 3) for the overall cohort and three clinical subgroups. The plots illustrate the class imbalance and prediction patterns across different clinical scenarios, with notably skewed distributions in adverse subgroups.

Groups	Accuracy	Kappa	F1(0)	F1(1–2)	F1(>=3)
All cohort	0.678	0.500	0.738	0.589	0.722
Age >= 35	0.696	0.472	0.790	0.545	0.673
Number of embryos <= 4	0.710	0.365	0.802	0.543	0.140
No good embryos	0.675	0.396	0.771	0.560	0.478

Table 3. Model performance metrics across the overall cohort and subgroups.

mapping approach fails to identify key factors and struggles to fully utilize features that have been shown to have independent effects in cycle-level binary predictions, such as the proportion of high-quality embryos and female age[8,9]. This methodological shift not only enhanced performance but also provides biological insights, which are significant for health workers in applications[23].

Regarding optimal model selection, various levels of assessment exist. First and foremost, accuracy is crucial; a method lacking accuracy is irrelevant, even if it is easy to understand[24]. Machine learning's superior performance highlights the limitations of traditional linear regression in capturing complex, non-linear interactions inherent to blastocyst development[25]. Moreover, the close performance among LightGBM, XGBoost, and SVM suggests that we may be approaching the theoretical prediction limit with the current feature set. Further improvements of predictive accuracy may require the identification of novel predictors.

For high-performing models, we then focus on their interpretability– a critical ethical consideration in IVF practice[21]. For clinicians, using decision support without the underlying factors driving those decisions is difficult. Fewer biomarkers in treatment guidelines enhance clinicians' comprehension, making the number

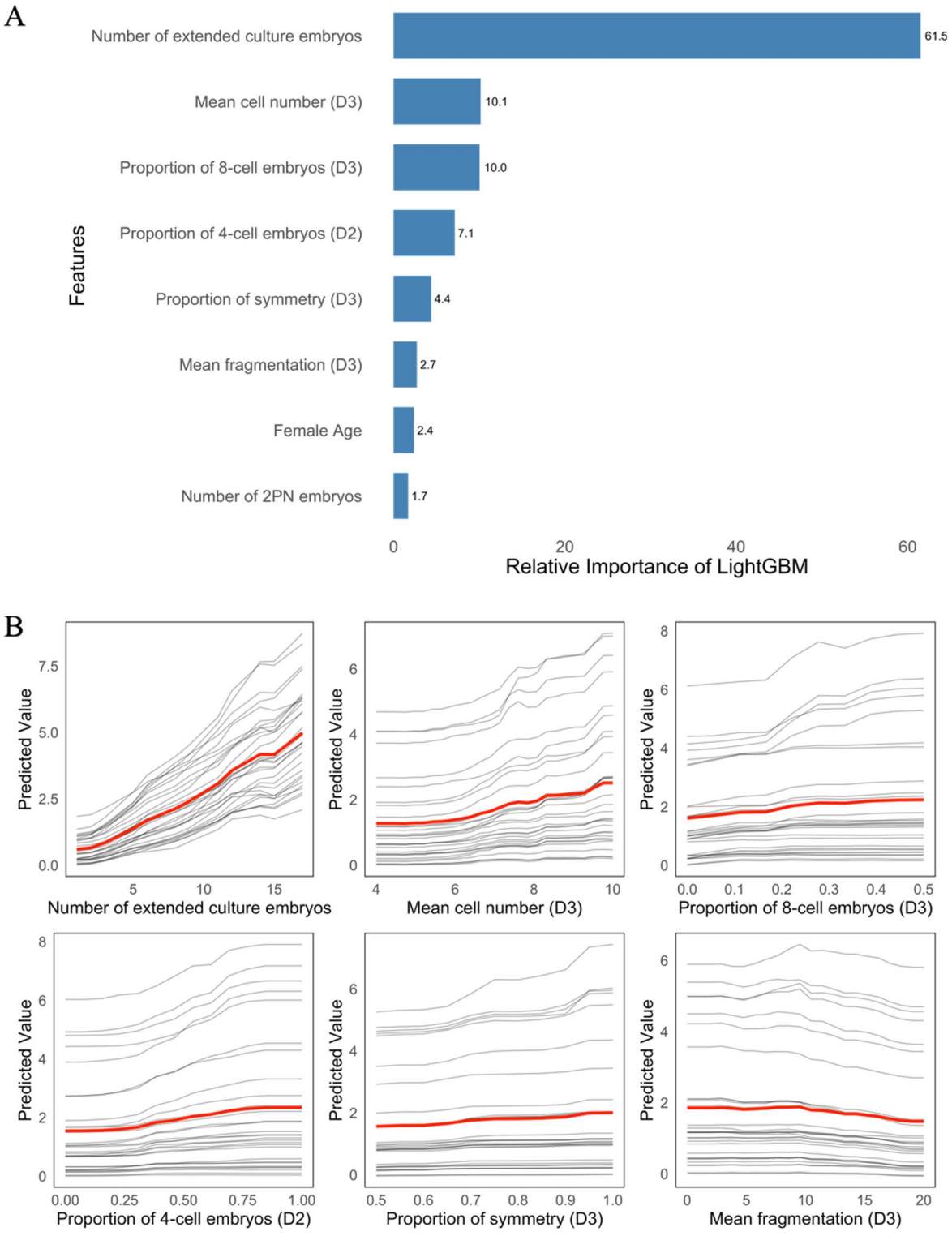

Fig. 3. Feature importance and partial dependence analysis using LightGBM. (**A**) The bar plot reveals the relative importance of features in the LightGBM model, with values quantifying each feature's proportional contribution to the model's predictive performance. (**B**) Individual conditional expectation and partial dependence plots illustrate the nuanced effects of the top six features on blastocyst yields. Thirty gray lines track the prediction trajectories of 30 samples, illustrating how predictions dynamically shift as a specific feature varies while other features remain constant. The red line delineates the mean effect across all samples, providing a comprehensive view of each feature's impact on model predictions.

of variables a key factor in assessing a method's interpretability[24,26]. The understandability of methods is also a key aspect of explainability[24]. While both tree-based models, such as LightGBM, and kernel-based models, like SVM, can visualize their decision-making processes, SVMs tend to be less intuitive and harder to comprehend raising concerns about their visualizations[27]. LightGBM, which utilizes fewer features and is relatively easy to understand, strikes a balance between performance and interpretability, which is supported by a recent study predicting the number of oocytes retrieved[28].

Our interpretability analysis provides a global view on the cohort using importance ranking and also instance-level explanations through ICE plots[26]. Importance ranking revealed the number of embryos for extended culture as the primary predictor of blastocyst yield. This highlights the importance of having an adequate number of embryos to optimize extended culture outcomes, a key focus of previous studies and guidelines aimed at avoiding extended culture failure[6,29]. However, other previously recognized predictors such as 2PN embryos[11] and oocyte yield[11,12]did not emerge as significant factors in our model. Their strong correlation with the number of embryos for extended culture (r: 0.76–0.84) suggests they may influence blastocyst yield indirectly, with this number acting as a mediating factor.

Interestingly, partial independency plots revealed that blastocyst yields progressively increased with higher mean cell numbers in Day 3 embryos, even beyond the eight-cell stage. This finding highlights the developmental advantage of rapid-cleaving embryos during in vitro culture, consistent with prior studies[30,31]and warrants further investigation. Additionally, the proportion of four-cell embryos on Day 2 emerged as an independent predictor, aligning with recent research[32,33]. These findings suggest that current assessment protocols may undervalue early developmental kinetics[34].

ICE plots provide natural direct effects by holding confounders other than exposures constant to isolate the effects caused by changes in exposure[35]. We observed that single-feature interventions elicit varied responses across individuals. For instance, increasing the number of embryos for extended culture to five resulted in predicted blastocyst yields ranging from 0 to 3 across cycles. Such variability underscores the importance of personalized interventions in clinical practice, moving beyond single-feature threshold approaches. Understanding the complex interplay of multiple factors is crucial for guiding future policy development and research directions.

The subgroup analysis revealed variations in our predictive model for clinically complex cases. While accuracy remained stable (68-71%), predictive consistency declined (kappa: 0.37–0.50). In the "low embryo numbers" subgroup, the model performed poorly in predicting cycles with ≥3 blastocysts (F1 score = 0.14) but achieved a high F1 score (0.8) in the 0-blastocyst category. This aligns with Bayes' theory, which indicates that false alarm rates are high for events with low baseline occurrence rates (≥3 blastocysts account for only 2.5%, while 0 blastocysts account for 61.6%). Overall, due to the severe class imbalance in the subgroups, the model primarily demonstrates binary classification capabilities and struggles with nuanced high or low yield predictions. However, since having ≥3 blastocysts in subgroups is a clinically rare event, this does not undermine the model's utility.

This study also has several limitations. The single-center retrospective design and the lack of evaluation on an independent external cohort may constrain the robustness of the model. Furthermore, the non-all-blastocyst-culture policy at our center introduced an inherent selection bias: by transferring or freezing the highest-quality embryos on Day 3, the remaining embryos available for extended culture were likely of lower developmental potential. This practice may have led to an underestimation of actual blastocyst yield compared to settings that pursue whole-cohort blastocyst culture. As a result, our dataset reflects only a subset of clinical workflows 36[38]. Future prospective studies including all embryos produced would help develop more comprehensive and generalizable prediction models.

Despite these limitations, our study represents an important step toward personalized reproductive medicine by providing a quantitative framework for predicting blastocyst yields and identifying key biological determinants. Note that our framework is intended as a proof-of-concept, demonstrating that quantitative blastocyst yield prediction is feasible through tailored model development based on local clinical practices. By supporting individualized clinical decision-making and patient counseling, this approach may help reduce psychological stress associated with uncertainty in IVF outcomes. Moving forward, efforts to validate and refine this model across diverse clinical settings, including centers employing full-cohort blastocyst culture or using time-lapse technologies, will be essential. Additionally, integrating dynamic morphokinetic data and molecular biomarkers into predictive models could further enhance their accuracy and clinical applicability.

Methods

Study subjects

A retrospective study was conducted at the Center for Reproductive Medicine, Nanfang Hospital, Southern Medical University, China, between January 2016 and May 2022. The study was approved by the Institutional Review Board of Nanfang Hospital, as authorized by the Ethical Committee (approval number: NFEC-2024-326). Informed consent was waived due to the use of non-identifiable patient records.

The inclusion criteria were as follows: (1) Autologous IVF/ICSI cycles performed at our center without egg or sperm donation; (2) Cycles with embryos undergoing extended culture to day 5/6; (3) No restrictions on patient characteristics (e.g., female age), ovarian stimulation protocols (conventional antagonist, agonist, or other unconventional regimens), and gamete states (fresh or frozen). The exclusion criteria included: (1) Cycles recorded with missing blastulation outcomes; (2) Cycles that contained an exceptionally high number of embryos in extended culture (> 20). We initially collected a total of 9,857 blastocyst culture cycles. After applying exclusion criteria, we removed 128 cycles with incomplete outcome labels and 80 cycles with an unusually high number of embryos. Our final dataset included 9,649 cycles.

Treatment procedures

Most patients underwent personalized ovarian stimulation protocols using standard agonist or antagonist regimens, while a few received alternative stimulation approaches, such as minimal stimulation, individualized based on patient-specific characteristics including age, antral follicle count, menstrual cycle, body mass index (BMI), and anticipated ovarian response. Ovarian stimulation typically began between the 2nd and 5th days of the menstrual cycle or during the luteal phase. Patients received 75–375 IU of gonadotropin daily during the ovarian stimulation process. Recombinant FSH or urinary hMG (human menopausal gonadotropin) was administered for gonadotropin stimulation. The initial dosage was maintained for 4 days and then adjusted based on the patient's follicular growth response and serum E2 levels. For antagonist protocols, GnRH antagonists were flexibly administered until the trigger day. Final oocyte maturation was triggered when at least one follicle reached 18 mm in diameter, using intramuscular human chorionic gonadotropin (hCG: 2000–10000 IU, Livzen, China; or 250 μg, Ovidrel, Merck-Serono, Switzerland), with or without Triptorelin (0.2 mg, Decapeptyl, Ferring, Switzerland). Oocyte retrieval occurred 34–36 h post-hCG administration under transvaginal ultrasound guidance.

Both conventional IVF and ICSI were employed as primary fertilization techniques. For a few patients with severe male factor infertility, percutaneous epididymal sperm aspiration or testicular sperm aspiration were additionally utilized to retrieve sperm for subsequent fertilization. Following insemination, oocytes were cultured in pre-equilibrated cleavage medium under mineral oil in incubators maintained at 37 °C, 6% CO_2, and 5% O_2 in a humidified atmosphere. On Day 3, the top embryos (with 73% being the top 1–2) was selected for fresh transfer or cryopreservation and the remaining embryos were extended cultured in blastocyst medium under the same conditions. This non-all-blastocyst-culture strategy aims to avoid having no embryos for transfer after a failed extended culture. Embryo development was monitored according to the Istanbul consensus timelines[34]with detailed records of cell number, fragmentation, and blastomere symmetry on Day 2 and 3. According to the Istanbul consensus, a good day-3 embryo was defined as having 8 equally-sized, mononucleated blastomeres with less than 10% fragmentation. Blastocyst formation was defined by the development of viable embryos suitable for transfer or cryopreservation. Usable blastocysts were classified as those embryos achieving a Gardner score of ≥ 3BC on Day 5 or 6. Blastocyst yield was calculated as the total number of usable blastocysts per cycle.

Feature selection

A comprehensive set of 21 features were selected from the database based on three criteria: (1) temporal relevance – only data collected before extended culture initiation were included; (2) literature-based predictors – twelve established factors such as female age, fertilization method, number of extended culture embryos[8]infertility cause, number of oocytes retrieved, and number of normally fertilized oocytes (2 pronuclei;[11] were incorporated. Additionally, Day 3 embryo properties (e.g. proportion of 8-cell and > 8-cell embryos, mean cell number, proportion of < 10% fragmentation rate, mean fragmentation and proportion of symmetry) were included[11,19]; (3) expert-driven parameters – nine clinically relevant features such as patient demographics (male age, BMI, infertility type), stimulation protocols, total gonadotropin dosage, stimulation duration, and Day 2 embryo characteristics (e.g. proportion of 4-cell embryos, proportion of > 4-cell embryos and mean cell number) were selected.

Missing data were addressed through imputation methods, with mean imputation applied to continuous variables (e.g., female age, BMI) and mode imputation for categorical variables (e.g., fertilization method, stimulation protocol). All included variables demonstrated missing rates below 5%. Detailed definitions of the included features are provided in Supplementary Table S1, and Pearson correlation coefficients among numerical variables are shown in Supplementary Figure S1.

Recursive Feature Elimination (RFE), a wrapper-based feature selection method[39]was applied to enhance model performance and reduce dimensionality. RFE was conducted independently for each model, beginning with the full set of features. Feature importance was iteratively assessed, and the least significant features were systematically removed until only two features remained. Performance metrics for each model were recorded at each iteration across feature subsets of varying sizes. The optimal feature subset was identified by balancing model performance and feature parsimony. The highest R-squared (R^2) value achieved across all feature subsets was first identified, and a tolerance threshold of 0.005 below this maximum was applied. Among all the feature subsets meeting this criterion, the one containing the fewest features was selected as optimal. This approach ensured the selection of a minimal feature set while maintaining near-optimal performance.

Model development

For model development, the dataset was randomly split into training (70%) and test (30%) sets, with the test set held out during training to ensure unbiased performance evaluation. Three machine learning models – Support Vector Machine (SVM), Light Gradient Boosting Machine (LightGBM) and Xtreme Gradient Boosting (XGBoost) – were evaluated alongside a linear regression baseline to capture linear and non-linear relationships, handle feature interactions, and mitigate overfitting.

Hyperparameter optimization was performed using grid search with 5-fold cross-validation[40]. The training set was divided into five equal parts, with four used for training and one for validation in each iteration. This process was repeated five times, allowing each subset to serve as validation data once. Performance metrics were averaged across all iterations, providing reliable hyperparameter evaluation and reduced overfitting risks. After each feature reduction in RFE, optimal hyperparameters were re-searched to ensure the best model performance with the updated feature set. The optimal hyperparameter configurations for each model, along with the corresponding optimal feature set, are detailed in Supplementary Table 2. Following optimization, each model was retrained on the full training set using the identified optimal hyperparameters and feature subset. Model performance was then assessed on the held-out test set to evaluate the generalization capacity.

Model evaluation

All analyses were performed using R Statistical Software (v4.4.1, RStudio 2023.9.1). Regression performance was measured using R^2, mean absolute error (MAE), and root mean square error (RMSE). R^2 quantifies the proportion of variance in the dependent variable explained by the model, with values ranging from 0 to 1, where higher values indicate better model performance. MAE represents the average absolute difference between predicted and actual values, with lower MAE indicating greater accuracy. RMSE, calculated as the square root of the mean squared differences between predicted and actual values, penalizes larger errors more heavily than MAE, making it more sensitive to outliers.

After consulting two experienced clinicians, it was determined that considering blastocyst yield in ranges, rather than as precise numbers, holds greater practical significance. Consequently, the predict values of the target variable were discretized into bins, resulting in a three-category system (0, 1–2, and ≥ 3 blastocysts). The three-class classification was evaluated using multiple metrics, including accuracy, Cohen's kappa coefficient, and F1 scores. Accuracy reflects the proportion of correct predictions out of the total predictions made, with values ranging from 0 to 1, where higher values indicating better performance. However, as accuracy can be misleading in cases of imbalanced class distributions, Cohen's kappa coefficient and class-specific F1 scores were also employed to provide a more comprehensive evaluation. Cohen's kappa measures the level of agreement between predicted and actual categories while a numbering for the possibility of agreement occurring by chance. Kappa values range from −1 to 1 and are interpreted as follows: >0.80 as excellent, 0.61–0.80 good, 0.41–0.60 moderate, 0.21–0.40 fair, 0-0.20 poor, and < 0 worse than random. The F1 score, ranging from 0 to 1, reflects the balance between precision and recall. It provides a harmonic mean of these two metrics, offering insight into the model's performance across different classes. F1 scores above 0.9 are considered excellent, those above 0.8 very good, and scores exceeding 0.7 are deemed good. Class-specific F1 scores were calculated to assess performance across individual categories.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 7 March 2025; Accepted: 11 June 2025
Published online: 02 July 2025

References

1. Glujovsky, D. et al. Cleavage-stage versus blastocyst-stage embryo transfer in assisted reproductive technology. *Cochrane Database Syst. Rev.* **5**, Cd002118 (2022).
2. ASRM. Blastocyst culture and transfer in clinically assisted reproduction: a committee opinion. *Fertil. Steril.* **110**, 1246–1252 (2018).
3. Smeltzer, S., Acharya, K., Truong, T., Pieper, C. & Muasher, S. Clinical pregnancy and live birth increase significantly with every additional blastocyst up to five and decline after that: an analysis of 16,666 first fresh single-blastocyst transfers from the society for assisted reproductive technology registry. *Fertil. Steril.* **112**, 866–873e1 (2019).
4. Xiong, F. et al. Association between the number of top-quality blastocysts and live births after single blastocyst transfer in the first fresh or vitrified-warmed IVF/ICSI cycle. *Reprod. Biomed. Online.* **40**, 530–537 (2020).
5. Cornelisse, S. et al. Cumulative live birth rate of a blastocyst versus cleavage stage embryo transfer policy during in vitro fertilisation in women with a good prognosis: multicentre randomised controlled trial. *Bmj* **386**, e080133 (2024).
6. National Collaborating Centre for Ws & Children's H, National Institute for Health and Clinical Excellence: Guidance, in Fertility: Assessment and Treatment for People with Fertility Problems. Royal College of Obstetricians & Gynaecologists Copyright © 2013, National Collaborating Centre for Women's and Children's Health.: London. (2013).
7. Gardner, D. K. et al. A prospective randomized trial of blastocyst culture and transfer in in-vitro fertilization. *Hum. Reprod.* **13**, 3434–3440 (1998).
8. Dessolle, L. et al. A cycle-based model to predict blastocyst transfer cancellation. *Hum. Reprod.* **25**, 598–604 (2010).
9. Spies, N. C., Pisters, E. E. A., Ball, A. E., Jungheim, E. S. & Riley, J. K. A machine learning approach to predict blastocyst formation in vitro. *Fertil. Steril.* **111**, e47 (2019).
10. Wang, X., Dong, C. Y., Zhang, C. L. & Zhang, S. D. A cycle-based model to predict no usable blastocyst formation following cycles of in vitro fertilization in patients with normal ovarian reserve. *Reprod. Biol. Endocrinol.* **23**, 11 (2025).
11. Jones, G. M., Trounson, A. O., Lolatgis, N. & Wood, C. Factors affecting the success of human blastocyst development and pregnancy following in vitro fertilization and embryo transfer. *Fertil. Steril.* **70**, 1022–1029 (1998).
12. Stone, B. A., March, C. M., Ringler, G. E., Baek, K. J. & Marrs, R. P. Casting for determinants of blastocyst yield and of rates of implantation and of pregnancy after blastocyst transfers. *Fertil. Steril.* **102**, 1055–1064 (2014).
13. Curchoe, C. L. & Bormann, C. L. Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. *J. Assist. Reprod. Genet.* **36**, 591–600 (2019).
14. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *Bmj* **385**, e078378 (2024).
15. Liao, Q. et al. Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring. *Commun. Biol.* **4**, 415 (2021).
16. Motato, Y. et al. Morphokinetic analysis and embryonic prediction for blastocyst formation through an integrated time-lapse system. *Fertil. Steril.***105**: 376–84.e9 (2016).
17. Segal, T. R. et al. Development of a decision tool to predict blastocyst formation. *Fertil. Steril.* **109**, e49–e50 (2018).
18. Lee, S., Choi, K. H., Hwang, S. & Kim, J. Blastocyst formation prediction based on deep learning model from 3-Day embryo images in Time-Lapse incubator using data augmentation. *Fertility Reprod.* **05**, 701–701 (2023).
19. Jiang, X. et al. Does conventional morphological evaluation still play a role in predicting blastocyst formation? *Reprod. Biol. Endocrinol.* **20**, 68 (2022).
20. Lee, T., Natalwala, J., Chapple, V. & Liu, Y. A brief history of artificial intelligence embryo selection: from black-box to glass-box. *Hum. Reprod.* **39**, 285–292 (2024).
21. Riegler, M. A. et al. Artificial intelligence in the fertility clinic: status, pitfalls and possibilities. *Hum. Reprod.* **36**, 2429–2442 (2021).
22. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry.* **77**, 534–540 (2020).

23. Lee, T., Natalwala, J., Chapple, V. & Liu, Y. A brief history of artificial intelligence embryo selection: from black-box to glass-box. *Hum. Reprod.* (2023).
24. Gimeno, M., Sada Del Real, K. & Rubio, A. Precision oncology: a review to assess interpretability in several explainable methods. *Brief Bioinform*. 24. (2023).
25. Barnett-Itzhaki, Z. et al. Machine learning versus classic statistics for the prediction of IVF outcomes. *J. Assist. Reprod. Genet.* **37**, 2405–2412 (2020).
26. Papoutsoglou, G. et al. Machine learning approaches in Microbiome research: challenges and best practices. *Front. Microbiol.* **14**, 1261889 (2023).
27. Barredo Arrieta, A. et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *arXiv e-prints* (2019).
28. Ferrand, T. et al. Predicting the number of oocytes retrieved from controlled ovarian hyperstimulation with machine learning. *Hum. Reprod.* **38**, 1918–1926 (2023).
29. Xue, X., Li, W. & Li, M. Optimal number of high-quality cleavage-stage embryos for extended culture to blastocyst-stage for transfer in women 38 years and older. *Gynecol. Endocrinol.* **39**, 2181642 (2023).
30. Kong, X. et al. The relationship between cell number, division behavior and developmental potential of cleavage stage human embryos: A Time-Lapse study. *PLoS One*. **11**, e0153697 (2016).
31. Liu, J. et al. Developmental potential of different embryos on day 3: a retrospective study. *J. Obstet. Gynaecol.* **42**, 3322–3327 (2022).
32. Mi, Z. et al. Number of blastomeres in Day-2 embryos affect the rates of blastocyst formation and clinical pregnancy during in vitro fertilization cycles. *Reprod. Sci.* **28**, 3397–3405 (2021).
33. Feng, K. et al. Predictive factors for the formation of viable embryos in subfertile patients with diminished ovarian reserve: A clinical prediction study. *Reprod. Sci.* **31**, 1747–1756 (2024).
34. ASRM&ESHRE. The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. Hum Reprod. ; 26:1270-83. (2011).
35. Cox, L. A. Jr. Objective causal predictions from observational data. *Crit. Rev. Toxicol.* **54**, 895–924 (2024).
36. Wang, L., Qiu, P., Jiang, L., Li, P. & Jiang, Y. Competent blastocyst and receptivity endometrium improved clinical pregnancy in fresh embryo transfer cycles: a retrospective cohort study. *BMC Pregnancy Childbirth*. **24**, 258 (2024).
37. Wang, X., Xiao, Y., Sun, Z., Zhen, J. & Yu, Q. Strategy for embryo transfer to improve pregnancy outcomes in advanced maternal age. *Zygote* **30**, 781–789 (2022).
38. Wei, H., Zhu, B., Deng, L., Zeng, M. & Duan, J. Optimal embryo management strategies for patients undergoing antagonist protocols in IVF treatment. *J. Assist. Reprod. Genet.* **42**, 827–838 (2025).
39. Guyon, I., Weston, J. & Barnhill, S. Vapnik V. Gene selection for Cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
40. Bergstra, J. & Bengio, Y. Random search for Hyper-Parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).

Author contributions

WJ. H. conceptualized the research, performed the primary data analysis, and drafted the initial manuscript. F. P. contributed to the methodology development, statistical analysis, and critical revision of the data. S. Q. supervised the research design, provided critical scientific insights, and oversaw the research process. XC. W. conceived the original research idea, provided comprehensive guidance, and made substantial contributions to the manuscript's final version. All authors participated in data interpretation, critically reviewed the manuscript, and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China [82171656], Scientific and Technological Planning Project of Guangzhou City [2023A042302], and Nanfang Hospital [2023B035].

Declarations

Ethics approval and consent to participate

Ethical approval for the study was obtained from the Institutional Review Board of Nanfang Hospital, as authorized by the Ethical Committee (approval number: NFEC-2024-326). The procedures followed were in accordance with the ethical standards of the Declaration of Helsinki of the World Medical Association. The Ethical Committee of Nanfang Hospita waived the need for obtaining informed consent from the participants.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-06998-4.

Correspondence and requests for materials should be addressed to X.-c.W.