ORIGINAL RESEARCH PAPER

Prediction of in-vitro fertilization outcome by ultrasound strain analysis and machine learning: A multi-center study

ANYI CHENG[1,2,3*], YIZHOU HUANG[2*], CONNIE REES[2,4,5], CELINE BLANK[6,7], NIKOS CHRISTOFORIDIS[8], DICK SCHOOT[4,5], LIN XU[1,9], MASSIMO MISCHI[2]

*[1]School of Information Science and Technology, ShanghaiTech University, Shanghai, China; [2]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands; [3]Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China; [4]Department of Obstetrics and Gynecology, Catharina Hospital Eindhoven, Eindhoven, Netherlands; [5]Department of Reproductive Medicine, Ghent University Hospital, Ghent, Belgium; [6]Department of Reproductive Medicine, ZAS Augustinus, Antwerp, Belgium; [7]Department of Reproductive Medicine, University Hospital Leuven, Belgium; [8]Embryolab Fertility Clinic, Thessaloniki, Greece; [9]State Key Laboratory of Advanced Medical Materials and Devices, Shanghai, China. *Co-first authors: Anyi Cheng and Yizhou Huang*

ABSTRACT

Background: In-vitro-fertilization (IVF) failure rates remain above 65% with unknown causes. Uterine receptivity, largely determined by uterine peristalsis, is believed to play a key role in the IVF success. Accurate assessment of uterine peristalsis holds the potential for improving the success rate of embryo implantation.

Methods: This prospective study includes 62 IVF patients from multiple fertility centers under three different clinical settings. Four-minute B-mode transvaginal ultrasound (TVUS) scans were performed one hour before embryo transfer (ET). 25 features related to frequency, amplitude, power, velocity, and coordination were extracted using strain analysis from TVUS speckle tracking results. Three probabilistic classifiers, i.e., support vector machine (SVM), K-nearest neighbors (KNN), and adaptive boosting (AdaBoost), were employed to discriminate uterine activity as either favorable or adverse to clinical pregnancy rate. Prior to machine learning, feature selection was performed by categorized feature ranking and sequential forward selection. The proposed method was evaluated by a nested 8-fold cross validation.

Results: Our results suggest that features related to coordination and frequency of the uterine peristalsis are strongly associated with clinical pregnancy, and SVM demonstrates the best classification performance between successful and unsuccessful pregnancies, with an average area under the receiver operating characteristic curve of 0.81.

Received: 19 November 2025 | Accepted: 26 November 2025

Correspondence: Lin Xu / School of Information Science and Technology, ShanghaiTech University, Shanghai, China / E-mail address: xulin1@shanghaitech.edu.cn

Conclusions: We developed a machine learning framework to improve the prediction of IVF outcome based on TVUS recordings of patients from multiple centers. Our SVM model identified significant uterine motion features and demonstrated reliable and generalizable classification performance. This work can provide useful means to support clinicians for clinical decision-making prior to ET and possibly enhance IVF success rates.

Key words: Infertility, Transvaginal ultrasound, Uterine contractility, Speckle tracking, Machine learning

Introduction

In-vitro fertilization (IVF) is the most advanced clinical technology for infertility (1). Yet, its success rate stays around 35% (2). A typical IVF cycle involves stimulating follicles, retrieving oocytes, fertilizing them in vitro, and transferring an embryo into the uterus (embryo transfer, ET). The success of ET and subsequent implantation are crucial for IVF success, but representing stages of treatment with minimal control. Previous studies have employed machine learning to predict the success of ET, considering various clinical factors, such as demographic information, female/male pathology, semen quality, oocyte stimulation response, and embryological data (3-6). Additionally, uterine peristalsis, which refers to rhythmic uterine contractions mostly visible under the endometrium, may influence IVF success (7-9). Low or absent uterine contractility is preferred during the embryo implantation phase (10). Pharmacological intervention or surgical corrections to alter uterine contractions can enhance implantation success (11-13). A comprehensive characterization of uterine contractions outside pregnancy is essential to identify the patients who may benefit (most) from regulating uterine contractility (14). Transvaginal ultrasound (TVUS) offers a non-invasive and cost-effective method to measure uterine activity (15). Quantitative strain imaging performed by speckle tracking techniques, such as block matching (BM) (16-18) and optical flow (OF) (19, 20), can analyze tissue displacement throughout the TVUS recording, and allow for a good and objective description of uterine deformation based on strain analysis (17). A recent pilot study combined electrohysterography (EHG) and TVUS to predict uterine activity as favorable or adverse to successful ET using machine learning (18). However, EHG features demonstrated limited predictive value compared to TVUS features. Studies have shown that embryo transfer outcomes are more directly influenced by uterine peristalsis in the subendometrial layer, which TVUS can assess with high resolution, whereas EHG primarily captures electrical activity in the outer myometrial layer (7-9). Furthermore, combining EHG and TVUS introduces complex measurement setup and signal processing, which hinders clinical adoption. Developing predictive models based on TVUS alone can simplify the protocol and accurately characterize uterine peristalsis, potentially leading to improved clinical decision-making, e.g., postponing ET to a natural cycle or after pharmacological modulation of the uterine activity (13, 21). The aim of the present study is to build a machine learning framework to enhance the generalizability and reliability of IVF outcome predictions (prior to ET) based on quantitative analysis of uterine motion measured by TVUS only.

Methods

Study design

This multi-center observational prospective study was approved by the respective local ethics committees and all included patients provided written informed consent. All three clinical settings included patients undergoing IVF/Intracytoplasmic Sperm Injection (ICSI) treatment, with fresh transfer of a single good or excellent quality embryo graded by the Istanbul conference Alpha criteria (22). Exclusion criteria included poor or moderate quality embryos, patients with severe (III/IV) endometriosis or adenomyosis,

hydrosalpinges and/or uterine anomalies. Each patient underwent a 4-minute B-mode TVUS recording of the uterus in mid-sagittal section within one hour before ET. 62 TVUS recordings were included in total.

Participants

From the IMPLANT 1 study (23), 31 IVF patients receiving placebo from one of the 60 participating fertility centers were included for secondary analysis in this study. These patients had a history of no more than one failed IVF cycle and underwent a gonadotropin-releasing hormone (GnRH) antagonist protocol with a single injection of human chorionic gonadotropin (HCG) as the ovulation trigger. ET was performed on day 5. As a multi-center study, these recordings were acquired using different brands of ultrasound scanners and probes. More details on the patient enrollment and IVF protocol can be found in (23). At Ghent University Hospital (Belgium), 14 IVF patients received GnRH for follicle stimulation and HCG injection trigger 34–36 hours before oocyte retrieval. ET was carried out on day 5. Eleven patients received an agonist protocol and three patients an antagonist protocol. TVUS recordings were acquired using a Samsung-Medison WS80A scanner equipped with a V5-9 transvaginal convex 4D probe. More details can be found in (4). The third group of patients was included from Embryolab Fertility Clinic (Thessaloniki, Greece), 17 women underwent IVF treatment using a short GnRH agonist protocol. ET was conducted on day 5. TVUS recordings were acquired using a GE Voluson SWIFT scanner equipped with a RIC5-9A-RS transvagianl convex 4D probe. The acquisition frame rate ranged from 25 to 30 frames per second. All 62 IVF-ET patients were divided into successful and unsuccessful ongoing pregnancy groups (Table 1). Successful ongoing pregnancy was defined as a positive fetal heart rate detected at a gestational age of at least 11 weeks (18).

Speckle tracking and feature extraction

TVUS speckle tracking is susceptible to acoustic artifacts such as shadowing and reverberation that distort speckle patterns. A quality check on each 4-min recording was performed to avoid significant and persistent artifacts near the endometrium prior to the tracking process. One of the primary challenges in applying speckle tracking to TVUS recordings is out-of-plane (OOP) motion from subject or operator movement, as well as the possible occurrence of shadowing. To mitigate these effects, a two-step tracking algorithm was applied, as detailed in (19): (1) Endometrial midline tracking markers (TMs; blue dots in Figure 1) were manually placed and tracked by OF (24). The global (rigid) translation and rotation of the endometrium in each frame was estimated by tracking the line fitting these TMs. (2) A predefined grid of TMs along the endometrial walls (red dots in Figure 1) was updated per frame using these motion estimates, so as to maintain their position relative to the uterine morphology. OF tracking was then applied to these TMs only between consecutive frames, maintaining their anatomical position and effectively limiting the impact of OOP motion and shadowing, as possible tracking errors did not accumulate over consecutive frames. Furthermore, tracking quality was validated via Pearson correlation coefficient (PCC) between consecutive frames. A sharp decrease in PCC indicated the presence of strong artifacts. A PCC>0.8 was considered as indicating valid tracking. Across 62 patients, average PCCs of 0.9926 ± 0.0026 and 0.9121 ± 0.2724 were obtained from tracking the midline TMs and grid TMs, respectively, confirming consistently high tracking quality.

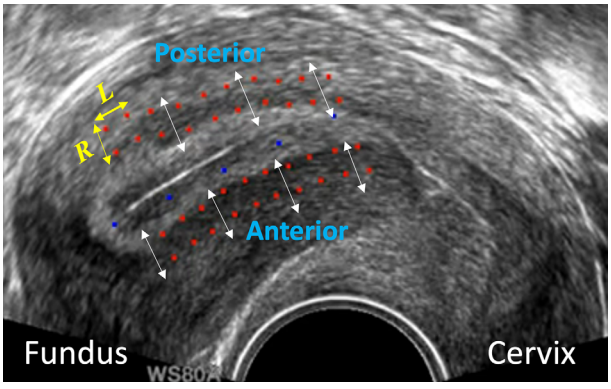

Figure 1. Grid positioning (red dots) alongside the endometrium for optical flow tracking in longitudinal (L) and radial (R) directions. 2D Transvaginal ultrasound image in midsagittal plane acquired from a healthy volunteer during the late follicular phase.

Table 1. Baseline characteristics of the study group.

Patient Characteristics	Ongoing Pregnancy (n= 26)	No Ongoing Pregnancy (n=36)	*p*-value[a]
Setting (n (%))			
• IMPLANT Study (n=34)	10 (38.5%)	24 (66.7%)	0.066[c]
• Ghent University Hospital (n=14)	7 (27.9%)	7 (19.4%)	
• Embryolab Fertility Clinic (n=14)	9 (34.6%)	5 (13.9%)	
Age (years)	35.50 ± 7.21	33.40 ± 5.38	0.198[d]
BMI (kg/m^2)	24.89 ± 3.43	24.63 ± 4.90	0.880[d]
Type of treatment (n (%))			
• IVF	3 (11.5%)	4 (11.1%)	0.687[c]
• ICSI	20 (76.9%)	30 (83.3%)	
• Both	3 (11.5%)	2 (5.5%)	
Infertility time (months)	34.71 ± 36.49	64.50 ± 35.02	0.165[d]
Cause of infertility (n (%))[b]			
• Ovulation disorder	5 (41.7%)	3 (27.3%)	0.747[c]
• Male factor	2 (25.0%)	4 (36.4%)	
• Tubal Factor	1 (8.3%)	2 (18.2%)	
• Idiopathic	2 (16.7%)	2 (18.2%)	
Number of previous IVF cycles	2.15 ± 2.48	2.51 ± 1.64	0.707[d]
Endometrial thickness at ET (mm)	11.14 ± 2.55	8.55 ± 2.41	**0.032**[d]
Stimulation protocol (n (%))			
• Agonist protocol	13 (50.0%)	12 (33.3%)	0.203[c]
• Antagonist protocol	13 (50.0%)	24 (66.7%)	
Gravidity	1.17 ± 1.53	1.18 ± 1.68	0.983[d]
Parity	0.25 ± 0.45	0.18 ± 0.40	0.708[d]

Data are presented as n (%) or mean ± standard deviation.[a] *p*-value was considered significant at <0.05.

[b] Cause of infertility details were missing for the IMPLANT study subgroup due to limitations to data access. [c] Chi-squared analysis. [d] T-test for independent samples. Abbreviations: BMI: body mass index; IVF: in vitro fertilization; ICSI: intracytoplasmic sperm injection; ET: embryo transfer.

Strain in radial and longitudinal directions, representing uterine muscle deformation, was calculated based on the displacements of each pair of adjacent tracking points. The study investigated 25 motion-related (Table 2) features derived from the strain signals in radial and longitudinal directions, originating from the anterior and posterior walls: contraction frequency (CF), which counts zero-crossing instances per minute in the strain signal (18); mean frequency (MF), the average frequency of strain signals in the Fourier domain; standard deviation (STD), representing the strain signal amplitude; and unnormalized first statistical moment (UFM), measuring power with amplitude and frequency. Additional features were based on spatiotemporal analysis of the ultrasound video loops. Next to the longitudinal velocities (V) of the uterine contractions in both cervix to fundus (C2F) and fundus to cervix (F2C) directions (19), the contraction coordination between the anterior and posterior walls, was assessed using five similarity measures between the propagation directions: correlation coefficient (CC) (19), mean squared error (MSE), Hausdorff distance (HD) (19), Euclidean disatance (ED), and concordance correlation coefficient (CCC) (25).

In our previous study, we evaluated reproducibility and variability of velocity and coordination features in ten healthy volunteers. The intraobserver reproducibility test was performed using three repeated grid positionings by one engineer, yielding excellent agreement (intraclass correlation coefficient, ICC > 0.9). For the interobserver variability assessment, two clinicians and an engineer independently analyzed the data, and good agreement was obtained (ICC > 0.75) (19).

Feature selection

The extracted features were fed into a dedicated machine learning framework to predict the IVF outcome, as depicted in Figure 2. Feature selection was performed on the training data in the inner cross validation (CV) loop (Figure 2b) to identify a subset of relevant features from the original feature pool, mitigating overfitting problems and leading to faster and more accurate model realization. Prior to feature selection, feature normalization was implemented on the training data to reduce the influence of different feature scales on the machine learning algorithms. A categorized feature ranking approach (Figure 2c) was first implemented by selecting features from four categories: frequency, amplitude and power, velocity, and coordination. Wilcoxon rank sum test was employed to assess the statistical difference in every feature between successful and unsuccessful groups (26). All the features were ranked within each category according to their p-values in ascending order. The first N features in each category were chosen under the assumption that a lower p-value indicates better classification ability. The optimization of N was determined by a search within {2,3,4}. The optimum value was chosen based on the accuracy on the inner validation set. Sequential forward selection (SFS) was applied to the selected features for searching the optimal feature combination (27). The SFS procedure started with an empty set and sequentially loaded one feature at each iteration. Additional features were then incrementally incorporated if their inclusion alongside existing features produced

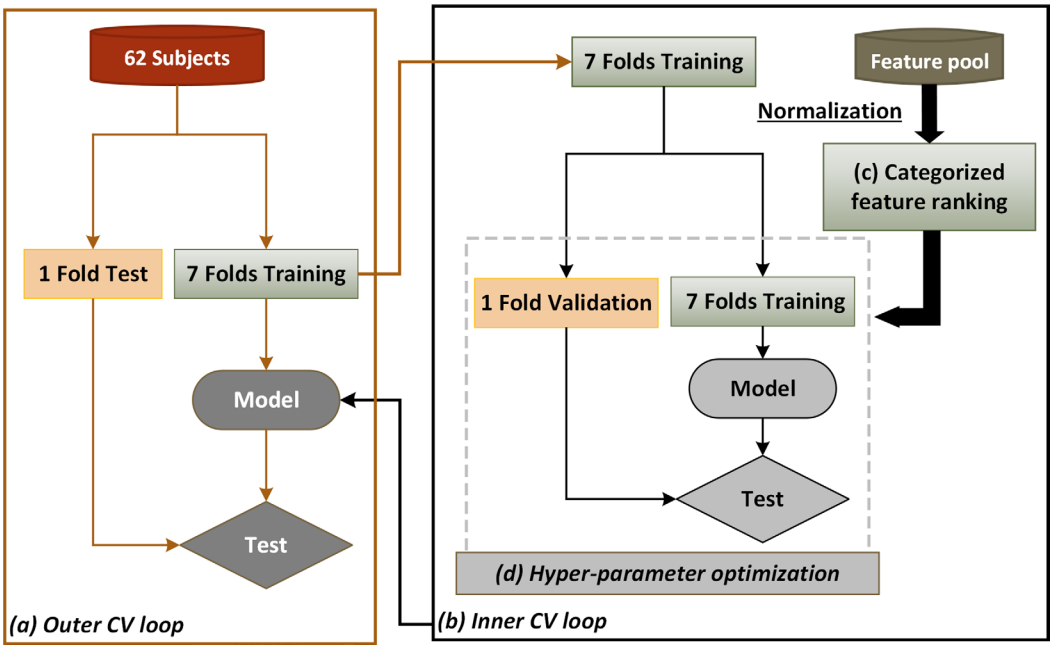

Figure 2. Schematic of the nested cross-validation (CV) framework. (a) Outer CV Loop: The 62-patient dataset is split into 8 folds, with one-fold (7–8 patients) reserved as the test set and seven folds (54–55 patients) used as the training set. (b) Inner CV Loop: the training set is further partitioned using stratification into 7 folds (6-7 patients each) and 1 fold (6-7 patients). (c) Categorized feature ranking: Features are ranked and filtered within each category by Wilcoxon rank sum test. (d) Hyper-parameter optimization for classifiers. The entire nested CV procedure is executed 10 times with randomized partitions.

the most substantial enhancement. This process continued until adding a new feature no longer yielded further improvements in accuracy.

Nested cross-validation

A nested CV procedure, comprising an inner CV loop (Figure 2b) nested within an outer CV loop (Figure 2a), was employed to develop a conservative and generalizable classification model (Figure 2). The 62-patient dataset was first partitioned into 8 folds in a stratified manner to ensure the proportion of each class (i.e., successful vs. unsuccessful pregnancy, 26:36) is preserved in each fold. The training set consisted of 7 folds with 7-8 patients each for model development, while the test set contained the remaining 1 fold (7-8 patients), strictly reserved for final evaluation. In the inner CV loop, the training set of 54-55 patients was further partitioned using stratification into 7 folds (6-7 patients each) and 1 fold (6-7 patients) to guide feature selection and hyperparameter tuning. Features were first filterd by categorized feature ranking and iteratively added by SFS until the optimal feature combination was identified, while hyperparameters were optimized to maximize validation accuracy. This inner loop ensures that the optimal feature subset and hyperparameters are identified without exposing the model to the outer test data. The finalized model with the selected features and tuned parameters was evaluated on the outer test fold. To reduce the uncertainty, the whole nested CV process was repeated 10 times with random splits of the observations into 8 folds in each repetition. The average model performance over the 10 repetitions was considered.

Classifiers

Three classifiers were trained and evaluated individually to classify successful and unsuccessful pregnancy groups. To address the slight class imbalance, we implemented class weights for all three classifiers in the inner training set. Class weights were set inversely proportional to class frequencies, assigning higher penalties to misclassifications of the minority class to enhance model sensitivity. Support vector machine (SVM), is a powerful classifier (28). The radial based function kernel was used to improve the efficiency of data separation (29). Hyper-parameter optimization (Figure 2d) was implemented using a full grid-search method (30). To avoid overfitting, the searching ranges adopted were limited to $\{2^{-2}, 2^{-1},...,2^{2}\}$ for both the cost and the kernel scale. In SVM, the prediction threshold is a critical parameter used to determine the decision boundary for classifying data points into different classes. The optimal value of the threshold was identified by searching in the range of 0.1 to 0.9 in steps of 0.1. The best hyper-parameters were selected according to the accuracy metric on the inner validation set. By assuming the similarity of observations, K-nearest neighbors (KNN) classifies an observation based on a majority vote of its K nearest neighbors. The optimal number of classes, K, was determined by a search in the inner CV loop within the range {1, 3, 5, 7} (18). The best K was selected according to the evaluation accuracy in the inner CV loop. Adaptive boosting (AdaBoost) generates a strong classifier from a set of weak learners (classifiers) by adjusting adaptively the weights of samples and classifiers (31, 32). The number of weak learners L was optimized in the inner CV loop in the present study in the range of 10 and 40 in steps of 5. The best L was chosen based on the accuracy metric of the inner loop.

Model evaluation

For each nested CV repetition, a confusion matrix was obtained from the test set in the outer loop, from which accuracy (ACC), sensitivity (SN), and specificity (SP) were derived to evaluate the performance of the machine learning models with selected features and hyper-parameters in the inner loop. To assess the classification power of each classifier to distinguish between successful and unsuccessful pregnancy, a receiver operating characteristic (ROC) curve for each classifier was obtained by plotting the true positive rate (TPR) vs. the false positive rate (FPR). For each ROC curve, the area under the curve (AUC) was computed as a quantitative performance metric. The whole nested CV process was repeated ten times, and the average values of all the performance metrics (ACC, SN, SP and AUC) from these repetitions were considered.

Results

Patient inclusion and baseline characteristics

26 out of 62 included patients (41.9%) achieved an ongoing pregnancy (Table 1). Age, BMI, type of treatment, stimulation protocol, duration of infertility, number of previous IVF cycles, gravidity, and parity did not differ statistically significantly between pregnant and non-pregnant groups ($p > 0.05$). Only endometrial thickness at time of ET was statistically significantly thicker in the pregnant group ($p = 0.032$).

Statistical feature analysis

An overview of 25 features is reported in Table 2. Successful and unsuccessful pregnancy groups were compared by Wilcoxon Rank Sum test (non-Gaussian distribution) and two-tailed Student's t-test (Gaussian distribution). Significant differences between the two groups were observed in MF (mean frequency), UFM (unnormalized first statistical moment), V (velocity), and most coordination features.

Representative features

Given the adopted nested CV strategy, 80 models with different optimum feature combinations were obtained for each classifier, i.e., 8 folds in the outer CV loop time 10 repetitions. Leveraging these results, we calculated the occurrence frequency for each feature over 80 optimal combinations. Figure 3 shows the top 5 frequently selected features for the three adopted classifiers. Despite variations among the different classifiers, coordination-related features are the most frequently and consistently chosen in the optimum feature subset.

Classification performance

Table 3 shows the classification performance of the three classifiers in terms of ACC, SN, SP, and AUC. The results indicate SVM to produce the best classification performance, with an ACC, SN, SP, and AUC of 0.77 ± 0.06, 0.80 ± 0.05, 0.75 ± 0.07, and 0.81 ± 0.04, respectively. Figure 4 displays the averaged ROC curves for the three classifiers over 10 repetitions of the nested 8-fold cross validation.

Discussion

In this study, we investigated quantitative uterine motion assessment for improved prediction of IVF outcomes. A categorized feature ranking and SFS was

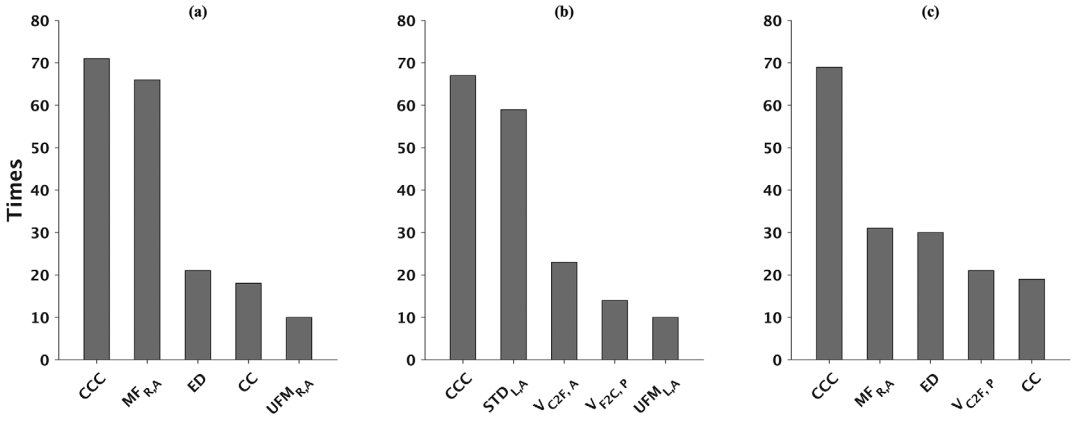

Figure 3. Top 5 frequently chosen features by (a) support vector machine (SVM), (b) adaptive boosting (AdaBoost), and (c) K-nearest neighbors (KNN). Whereas, 'A', 'P', 'R', and 'L' indicate features extracted from the anterior side, posterior side, in the radial and longitudinal direction, respectively. "C2F" and "F2C" represent propagation direction from cervix to fundus and from fundus to cervix, respectively. CCC: concordance correlation coefficient; CC: correlation coefficient; ED: Euclidean distance; UFM: unnormalized first statistical moment; MF: mean frequency; V: velocity; STD: standard deviation.

Table 2. Statistical feature analysis.

Category	Features	Ongoing Pregnancy (n= 26)	No Ongoing Pregnancy (n= 36)	*p*-value[a]	*AUC*[d]
Frequency (*UC/min*)	$CF_{L,A}$	1.40 ± 0.19	1.42 ± 0.18	0.65[b]	0.51
	$CF_{R,A}$	1.45 ± 0.20	1.54 ± 0.24	0.11[b]	0.67
	$CF_{L,P}$	1.41 ± 0.21	1.43 ± 0.17	0.70[b]	0.61
	$CF_{R,P}$	1.49 ± 0.27	1.56 ± 0.21	0.28[b]	0.51
	$MF_{L,A}$	1.39 ± 0.15	1.47 ± 0.16	0.09[c]	0.54
	$MF_{R,A}$	1.43 ± 0.23	1.52 ± 0.20	0.12[b]	0.66
	$MF_{L,P}$	1.43 ± 0.14	1.45 ± 0.14	0.62[b]	0.52
	$MF_{R,P}$	1.47 ± 0.19	1.57 ± 0.19	0.047[b]	0.58
Amplitude (%)	$STD_{L,A}$	7 (4)	5 (2)	0.07[c]	0.62
	$STD_{R,A}$	6 (5)	5 (3)	0.16[c]	0.62
	$STD_{L,P}$	6 (6)	6 (4)	0.95[c]	0.51
	$STD_{R,P}$	6 (4)	5 (3)	0.46[c]	0.55
Power (*Hz*)	$UFM_{L,A}$	1551.36 (1167.03)	893.23 (1058.75)	0.04[c]	0.60
	$UFM_{R,A}$	1116.20 (1341.03)	966.79 (835.46)	0.11[c]	0.59
	$UFM_{L,P}$	1086.46 (2766.33)	982.43 (1395.07)	0.63[c]	0.49
	$UFM_{R,P}$	1254.81 (1626.71)	849.63 (1017.33)	0.34[c]	0.55
Velocity (*mm/s*)	$V_{C2F,A}$	0.60 ± 0.16	0.65 ± 0.15	0.15[b]	0.59
	$V_{F2C,A}$	0.63 ± 0.17	0.68 ± 0.16	0.26[b]	0.60
	$V_{C2F,P}$	0.55 ± 0.17	0.64 ± 0.15	0.03[b]	0.67
	$V_{F2C,P}$	0.61 ± 0.15	0.65 ± 0.14	0.29[c]	0.60
Coordination (—)	*CC*	0.21 ± 0.31	-0.07 ± 0.27	0.0003[b]	0.77
	MSE	0.22 ± 0.09	0.34 ± 0.14	0.0002[c]	0.78
	HD	0.25 ± 0.12	0.27± 0.17	0.8697[c]	0.51
	ED	1.86 ± 0.36	2.29 ± 0.45	0.0002[c]	0.78
	CCC	0.17 ± 0.28	-0.06 ± 0.17	0.0003[c]	0.78

Data are presented as median (interquartile range) or mean ± standard deviation. [a]*p*-value was considered significant at <0.05. [b]T-test for Gaussian distributed data. [c]Wilcoxon-sum rank test for non-Gaussian distributed data. [d]AUC values were obtained from logistic regression analysis. Abbreviations: UC: uterine contractions; CF: contraction frequency; MF: mean frequency; STD: standard deviation; UFM: unnormalized first statistical moment; V: velocity; CC: correlation coefficient; MSE: mean square error; HD: Hausdorff distance; ED: Euclidean distance; CCC: Concordance correlation coefficient. 'A', 'P', 'R', and 'L' indicate features extracted from the anterior side, posterior side, in the radial and longitudinal direction, respectively. "C2F" and "F2C" represent propagation direction from cervix to fundus and from fundus to cervix, respectively.

Table 3. Performance metrics of three classifiers.

Classifier	ACC	SN	SP	AUC
SVM	0.77 ± 0.06	0.80 ± 0.05	0.75 ± 0.07	0.81 ± 0.04
AdaBoost	0.74 ± 0.04	0.65 ± 0.06	0.81 ± 0.05	0.80 ± 0.03
KNN	0.70 ± 0.05	0.57 ± 0.08	0.79 ± 0.09	0.74 ± 0.05

All data are presented as mean ± SD. Abbreviations: ACC: accuracy; AUC: area under the curve; SN: sensitivity; SP: specificity; SVM: support vector machine; AdaBoost: adaptative boosting; KNN: K-nearest neighbors.

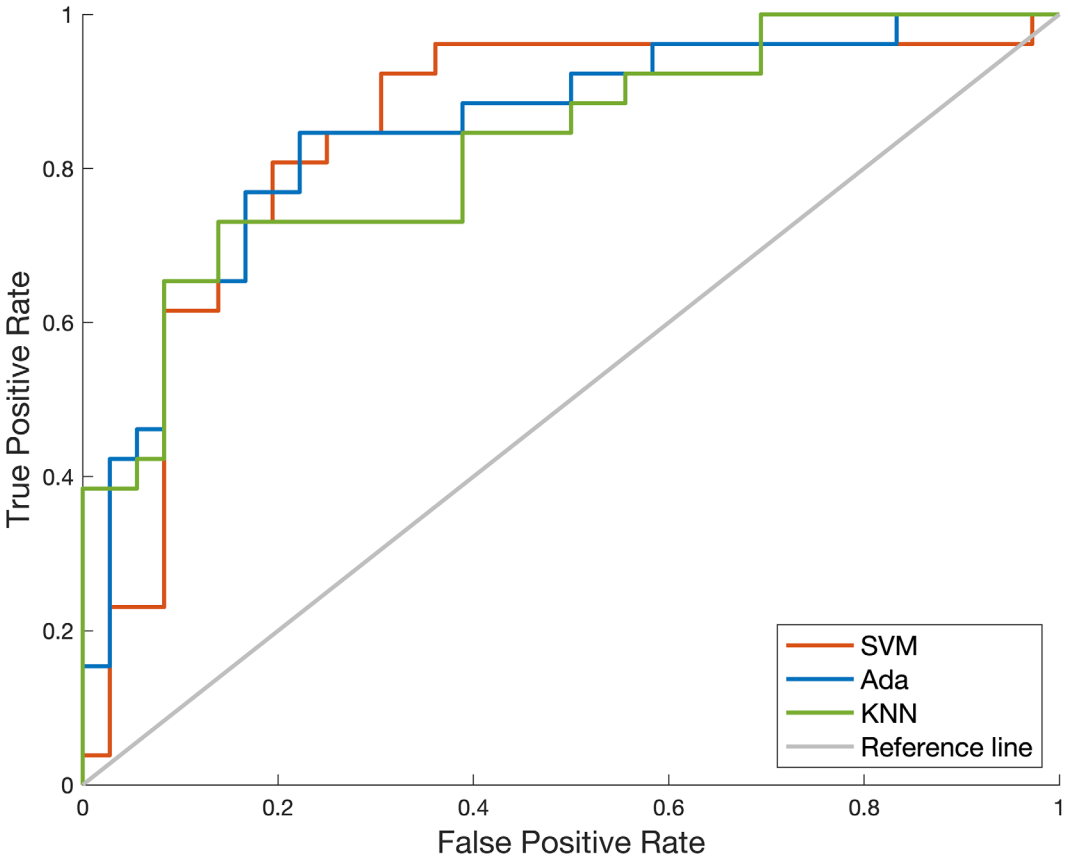

Figure 4. Averaged receiver operating characteristic (ROC) curves of the three classifiers. SVM: support vector machine; AdaBoost: adaptive boosting; KNN: K-nearest neighbors.

adopted to address the unique challenges of our limited dataset. Wilcoxon rank sum test identified significant features within each category, followed by SFS to optimize the remaining subset for improved classifier performance. This category-specific ranking ensured a balanced representation of uterine motion dynamics, allowing us to explore the combined effects of frequency, amplitude, velocity, and coordination features on IVF outcome prediction. By reducing the features to a subset of original, interpretable features that directly associated with relevant uterine motion patterns, we not only accelerated the optimization process, but reduced redundancy and overfitting risk as well as improved generalizability and clinical interpretability. While popular dimensionality reduction techniques like PCA or t-SNE are valuable for high-dimensional data, they transform the features into a new basis, which can obscure the physical meaning of the features, making it challenging to interpret results in the context of TVUS imaging. Our results (Figure 3) show that coordination features, describing the similarity of the uterine peristaltic waves alongside the endometrium, prevail in the top 5 features selected by all classifiers. This finding agrees with our statistical analysis in Table 2, where significant differences ($p < 0.05$) were found for most coordination features. In order to demonstrate the addictive value of our framework, we explicitly evaluate the classification performance of individual coordination metrics based on logistic regression analysis (Table 2). While coordination features (CCC: AUC 0.78; CC: AUC 0.77; ED: AUC 0.78) show strong individual performance, our SVM model based on feature selection achieves a 3–4% improvement in AUC. This aligns with our study's goal to explore and leverage mechanistic relationships between uterine activity and fertilization outcomes:

the obtained model performance gain suggests that no single metric fully explains the complexity of the uterine behavior, but their integration provides a more comprehensive view. Interestingly, MF was selected more than CF by all three classifiers. However, CF has previously shown better separation ability among four natural menstrual cycle phases (17) as well as between successful and unsuccessful pregnancies (18). This may be attributed to our feature selection procedure that the best individual features may not necessarily provide the most suitable information for boosting the model performance in combination with the other features. Note that, in this multi-center study, the TVUS recordings were collected using different scanners and probes, increasing significantly the inter-subject variability. Discrimination between successful and unsuccessful pregnancies is therefore more challenging than for single-center patients, as performed in (4, 5, 9, 18, 33). Despite this, the achieved average classification performance by SVM (ACC: 0.77, SN: 0.80, SP: 0.75, and AUC: 0.81) is still promising. Several machine learning models have attempted to predict IVF success based on clinical characteristics. Clustering-SVM was developed to predict cumulative pregnancy rate before starting the IVF with only patient information and obtained the AUC value of 0.7 (6). Similarly, Qju et al. (5) used pre-treatment clinical variables to estimate the cumulative live birth chance of the first IVF cycle. Extreme gradient boosting outperformed SVM and random forest (RF) on the validation set (AUC: 0.73) and provided an average ACC of 0.7 in the repeated nested CV using the whole dataset. Blank et al. (4) combined both embryonic and clinical variables for IVF implantation prediction and achieved better performance (AUC: 0.74) with RF. Much better performance (ACC: 0.94, SN: 0.86, SP: 1, and AUC: 0.86) with KNN was previously reported using uterine motion features extracted one hour before ET (18). Besides the differences in subject size (16 vs. 62 in our study) and origin, evaluation strategies between the two studies may also contribute significantly to the results. Although a nested CV strategy was used in (18), the hyper-parameter optimization was performed in the inner CV loop only, while feature selection and model evaluation were performed on the same test set. Such a CV strategy may cause a "data leakage" problem and lead to overestimated model performance (34). A stricter nested CV strategy was implemented in our study, where both feature selection and hyper-parameter optimization were performed in the inner CV loop. This ensures that the test set remains completely unseen during model training, serving exclusively for evaluating the performance of the trained model. This approach leads to a more reliable and generalizable estimation of the model performance. While this study identifies significant uterine motion features and achieves promising performance under strict validation, there are a few limitations that require further attention. Several works have proven the importance of clinical factors in IVF treatment success (3-6, 33), which were not evaluated in the present study. It maybe interesting to combine clinical and uterine motion features to have an extended characterization and perhaps further improve IVF outcome prediction. There is a significant number of missing data regarding patient characteristics for IVF indication in the IMPLANT study subgroup (Table 1). As we did not exclude patients according to IVF indications, this will likely not have any significant effect on the data or primary outcome, although the introduction of some confounding bias cannot be entirely excluded. Our study did exclude severe (III/IV) endometriosis or adenomyosis, hydrosalpinges and/or uterine anomalies patients showing aberrant contractility (35). Future studies could investigate if and how uterine contractility is specifically affected in these groups in the IVF context. Lastly, the limited size of the dataset can potentially lead to over-fitting and restrict the prediction generalizability of the models. To address this problem, the employed dedicated strategies, including class weighting on classifiers, feature selection, and nested cross-validation, were implemented to reduce complexity, prevent data leakage and overfitting, and promote generalizability. These measures were further validated by our multi-center dataset, which was collected using different scanner protocols and patient demographics. Importantly, this study represents a preliminary exploration aimed at identifying predictive uterine motion features and hyperparameters for IVF outcomes, rather than finalizing a fixed

model for clinical use. However, we acknowledge that external validation on a larger cohort is essential to confirm clinical utility. Future work will leverage the feature sets and hyperparameters identified in this preliminary study to train a streamlined, interpretable model optimized for clinical IVF applications. Under the significant success of deep learning in medical image analysis (36), training deep networks could also be considered when a larger dataset is available.

Conclusion

Our machine learning framework improves IVF outcome prediction for 62 patients across multiple fertility centers. Our findings highlight coordination and frequency features as crucial for embryo implantation. This indicates the potential use of TVUS signals to predict embryo implantation success, offering valuable insights for appropriate clinical decision-making and potentially increasing IVF success rates. Future research could integrate clinical and uterine motion features and employ deep learning on larger datasets.

List of abbreviations

IVF:	In-vitro fertilization
ET:	Embryo transfer
TVUS:	Transvaginal ultrasound
OF:	Optical flow
CV:	Cross validation
SFS:	Sequential forward selection
SVM:	Support vector machine
KNN:	K-nearest neighbors
AdaBoost:	Adaptive boosting

Ethics Approval and Consent to Participate: The study was approved by the Medical research Ethics Committee United (MEC-U) with reference number as A22.075/R15.043, issued on March 23rd, 2022. All included patients provided written informed consent.

Consent for Publication: All authors and editors consent to the publication of this data.

Availability of Data and Material: The patient data analyzed in the current study are not publicly available due to patient privacy.

Competing Interests: The authors declare no conflicts of interest regarding this manuscript.

Funding: This study has received funding both by GE Healthcare and the National Natural Science Foundation of China under Grant 62171284.

Authors' Contributions: AC, YH, DS, LX, and MM contributed to the design of the work. CR, CBI, and NC collected the data. YH and CR performed the interpretation of the data. AC and YH created the new software used in the work. AC has drafted the work. YH, CR, DS, LX, and MM have substantively revised the manuscript. All authors have approved the submitted manuscript.

References

1. Fauser BC. Towards the global coverage of a unified registry of IVF outcomes. Reprod Biomed Online 2019;38:133-7. https://doi.org/10.1016/j.rbmo.2018.12.001.
2. Andersen AN, Gianaroli L, Felberbaum R, De Mouzon J, Nygren KG. Assisted reproductive technology in Europe, 2002. Results generated from European registers by ESHRE. Hum Reprod 2006;21:1680-97. https://doi.org/10.1093/humrep/del075.
3. Raef B, Ferdousi R. A review of machine learning approaches in assisted reproductive technologies. Acta Inform Med 2019;27:205-11. https://doi.org/10.5455/aim.2019.27.205-211.
4. Blank C, Wildeboer RR, DeCroo I, Tilleman K, Weyers B, De Sutter P, et al. Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. Fertil Steril 2019;111:318-26. https://doi.org/10.1016/j.fertnstert.2018.10.030.
5. Qiu J, Li P, Dong M, Xin X, Tan J. Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method. J Transl Med 2019;17:1-8. https://doi.org/10.1186/s12967-019-2062-5.
6. Zhang B, Cui Y, Wang M, Li J, Jin L, Wu D. In vitro fertilization (IVF) cumulative pregnancy rate prediction from basic patient characteristics. IEEE Access 2019;7:130460-7. https://doi.org/10.1109/ACCESS.2019.2940588.
7. Zhu L, Che HS, Xiao L, Li YP. Uterine peristalsis before embryo transfer affects the chance of clinical pregnancy in fresh and frozen-thawed embryo transfer cycles. Hum Reprod 2014;29:1238-43. https://doi.org/10.1093/humrep/deu058.

8. Kuijsters NP, Methorst WG, Kortenhorst MS, Rabotti C, Mischi M, Schoot BC. Uterine peristalsis and fertility: current knowledge and future perspectives: a review and meta-analysis. Reprod Biomed Online 2017;35:50-71. https://doi.org/10.1016/j.rbmo.2017.03.019.
9. Blank C, Sammali F, Kuijsters N, Huang Y, Rabotti C, De Sutter P, et al. Assessment of uterine activity during IVF by quantitative ultrasound imaging: a pilot study. Reprod Biomed Online 2020;41:1045-53. https://doi.org/10.1016/j.rbmo.2020.08.006.
10. Tu Z, Ran H, Zhang S, Xia G, Wang B, Wang H. Molecular determinants of uterine receptivity. Int J Dev Biol 2014;58:147-54. https://doi.org/10.1387/ijdb.130345wh.
11. Pierzynski P, Reinheimer TM, Kuczynski W. Oxytocin antagonists may improve infertility treatment. Fertil Steril 2007;88:e19. https://doi.org/10.1016/j.fertnstert.2006.09.017.
12. Pierzynski P. Oxytocin and vasopressin V1A receptors as new therapeutic targets in assisted reproduction. Reprod Biomed Online 2011;22:9-16. https://doi.org/10.1016/j.rbmo.2010.09.015.
13. Moraloglu O, Tonguc E, Var T, Zeyrek T, Batıoglu S. Treatment with oxytocin antagonists before embryo transfer may increase implantation rates after IVF. Reprod Biomed Online 2010;21:338-43. https://doi.org/10.1016/j.rbmo.2010.04.009.
14. Rees CO, Thomas S, de Boer A, Huang Y, Zizolfi B, Foreste V, et al. Quantitative ultrasound measurement of uterine contractility in adenomyotic versus normal uteri: a multicenter prospective study. Fertil Steril 2024;121:864-72. https://doi.org/10.1016/j.fertnstert.2024.01.009.
15. Kim SH, Kim HD, Song YS, Kang SB, Lee HP. Detection of deep myometrial invasion in endometrial carcinoma: comparison of transvaginal ultrasound, CT, and MRI. J Comput Assist Tomogr 1995;19:766-72. https://doi.org/10.1097/00004728-199509000-00013.
16. Sammali F, Blank C, Xu L, Huang Y, Kuijsters NP, Schoot BC, et al. Experimental setup for objective evaluation of uterine motion analysis by ultrasound speckle tracking. Biomed Phys Eng Express 2018;4:035012. https://doi.org/10.1088/2057-1976/aab053.
17. Sammali F, Kuijsters NP, Huang Y, Blank C, Rabotti C, Schoot BC, et al. Dedicated ultrasound speckle tracking for quantitative analysis of uterine motion outside pregnancy. IEEE Trans Ultrason Ferroelectr Freq Control 2019;66:581-90. https://doi.org/10.1109/TUFFC.2018.2867098.
18. Sammali F, Blank C, Bakkes TG, Huang Y, Rabotti C, Schoot BC, et al. Multi-modal uterine-activity measurements for prediction of embryo implantation by machine learning. IEEE Access 2021;9:47096-111. https://doi.org/10.1109/ACCESS.2021.3067716.
19. Huang Y, Rees C, Sammali F, Blank C, Schoot D, Mischi M. Characterization of uterine peristaltic waves by ultrasound strain analysis. IEEE Trans Ultrason Ferroelectr Freq Control 2022;69:2050-60. https://doi.org/10.1109/TUFFC.2022.3165688.
20. Zhou Y, Zheng YP. A motion estimation refinement framework for real-time tissue axial strain estimation with freehand ultrasound. IEEE Trans Ultrason Ferroelectr Freq Control 2010;57:1943-51. https://doi.org/10.1109/TUFFC.2010.1642.
21. Kim SH, Riaposova L, Ahmed H, Pohl O, Chollet A, Gotteland JP, et al. Oxytocin receptor antagonists, atosiban and nolasiban, inhibit prostaglandin F2α-induced contractions and inflammatory responses in human myometrium. Sci Rep 2019;9:1-10. https://doi.org/10.1038/s41598-019-42181-2.
22. Balaban B, Brison D, Calderon G, Catt J, Conaghan J, Cowan L, et al. Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. Reprod Biomed Online 2011;22:632-46. http://doi.org/10.1016/j.rbmo.2011.02.001.
23. Griesinger G, Blockeel C, Pierzynski P, Tournaye H, Višňová H, Humberstone A, et al. Effect of the oxytocin receptor antagonist nolasiban on pregnancy rates in women undergoing embryo transfer following IVF: analysis of three randomised clinical trials. Hum Reprod 2021;36:1007-20. https://doi.org/10.1093/humrep/deaa369.
24. Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. In: Proc 7th Int Joint Conf Artif Intell; 1981; Vancouver, Canada. 2:674-9.
25. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255-68. https://doi.org/10.2307/2532051.
26. Wilcoxon F. Individual comparisons by ranking methods. In: Breakthroughs in statistics: methodology and distribution. New York: Springer; 1992. p. 196-202. https://doi.org/10.1007/978-1-4612-4380-9_16.
27. Marcano-Cedeño A, Quintanilla-Domínguez J, Cortina-Januchs MG, Andina D. Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In: Proc 36th Annu Conf IEEE Ind Electron Soc; 2010; Glendale, AZ, USA. p. 2845-50. https://doi.org/10.1109/IECON.2010.5675075.
28. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17. https://doi.org/10.1016/j.csbj.2014.11.005.
29. Hofmann T, Schölkopf B, Smola AJ. A review of kernel methods in machine learning. Max Planck Inst Tech Rep 2006;156.
30. Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. Dept Comput Sci, Natl Taiwan Univ 2003.
31. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55:119-39. https://doi.org/10.1006/jcss.1997.1504.
32. Li X, Wang L, Sung E. A study of AdaBoost with SVM based weak learners. In: Proc IEEE Int Joint Conf Neural Netw; 2005; Montreal, Canada. 1:196-201. https://doi.org/10.1109/IJCNN.2005.1555829.

33. Wang CW, Kuo CY, Chen CH, Hsieh YH, Su EC. Predicting clinical pregnancy using clinical features and machine learning algorithms in in vitro fertilization. PLoS One 2022;17:e0267554. https://doi.org/10.1371/journal.pone.0267554.
34. Tsamardinos I, Rakhshani A, Lagani V. Performance -estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. Int J Artif Intell Tools 2015;24:1540023. https://doi.org/10.1142/S0218213015400230.
35. de Boer A, Rees CO, Mischi M, Van Vliet H, Huirne J, Schoot BC. The influence of uterine abnormalities on uterine peristalsis in the non-pregnant uterus: a systematic review. J Endometr Uterine Disord 2023; 3: 100038. https://doi.org/10.1016/j.jeud.2023.100038.
36. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221-48. https://doi.org/10.1146/annurev-bioeng-071516-044442.