




A hybrid artificial intelligence model leverages multi-centric clinical data to improve fetal heart rate pregnancy prediction across time-lapse systems

A. Duval¹, D. Nogueira ^{2,3}, N. Dissler¹, M. Maskani Filali¹, F. Delestro Matos¹, L. Chansel-Debordeaux⁴, M. Ferrer-Buitrago ⁵, E. Ferrer⁵, V. Antequera⁵, M. Ruiz-Jorro⁵, A. Papaxanthos⁴, H. Ouchchane⁶, B. Keppi⁶, P.-Y. Prima⁷, G. Regnier-Vigouroux⁸, L. Trebesses⁹, C. Geoffroy-Siraudin¹⁰, S. Zaragoza¹¹, E. Scalici¹¹, P. Sanguinet⁸, N. Cassagnard², C. Ozanon¹², A. De La Fuente¹³, E. Gómez¹⁴, M. Gervoise Boyer¹⁰, P. Boyer¹⁰, E. Ricciarelli¹⁵, X. Pollet-Villard¹⁶, and A. Boussohier-Calleja ^{1*}

¹ImVtro, Paris, France ²INOVIE Fertilité, Institut de Fertilité La Croix Du Sud, Toulouse, France ³Art Fertility Clinics, IVF laboratory, Abu Dhabi, United Arab Emirate ⁴Service de la biologie et de la reproduction et CECOS, CHU Bordeaux Groupe Hospitalier Pellegrin, Bordeaux, France ⁵Crea Centro Médico de Fertilidad y Reproducción Asistida, Valencia, Spain ⁶INOVIE Fertilité, Gen-Bio, Clermont-Ferrand, France ⁷Laboratoire FIV PMAtlantique - Clinique Santé Atlantique, Nantes, France ⁸INOVIE Fertilité, LaboSud, Montpellier, France ⁹INOVIE Fertilité, AxBio, Bayonne, France ¹⁰Hopital Saint Joseph, Service Médecine et Biologie de la Reproduction, Marseille, France ¹¹INOVIE Fertilité, Bioaxiome, Avignon, France ¹²Clinique Hôtel Privé Natecia, Centre Assistance Médicale à la Procréation, Lyon, France ¹³Instituto Europeo de Fertilidad, Madrid, Spain ¹⁴Next Fertility, Murcia, Spain ¹⁵FIVMadrid, Madrid, Spain ¹⁶Nataliance, Centre Assistance Médicale à la Procréation, Saran, France

*Correspondence address. ImVtro, 130 rue Lourmel, 75015 Paris, France. E-mail: alexandra@im-vitro.com  <https://orcid.org/0000-0001-9524-454X>

Submitted on October 4, 2022; resubmitted on January 10, 2023; editorial decision on January 24, 2023

STUDY QUESTION: Can artificial intelligence (AI) algorithms developed to assist embryologists in evaluating embryo morphokinetics be enriched with multi-centric clinical data to better predict clinical pregnancy outcome?

SUMMARY ANSWER: Training algorithms on multi-centric clinical data significantly increased AUC compared to algorithms that only analyzed the time-lapse system (TLS) videos.

WHAT IS KNOWN ALREADY: Several AI-based algorithms have been developed to predict pregnancy, most of them based only on analysis of the time-lapse recording of embryo development. It remains unclear, however, whether considering numerous clinical features can improve the predictive performances of time-lapse based embryo evaluation.

STUDY DESIGN, SIZE, DURATION: A dataset of 9986 embryos (95.60% known clinical pregnancy outcome, 32.47% frozen transfers) from 5226 patients from 14 European fertility centers (in two countries) recorded with three different TLS was used to train and validate the algorithms. A total of 31 clinical factors were collected. A separate test set (447 videos) was used to compare performances between embryologists and the algorithm.

PARTICIPANTS/MATERIALS, SETTING, METHODS: Clinical pregnancy (defined as a pregnancy leading to a fetal heartbeat) outcome was first predicted using a 3D convolutional neural network that analyzed videos of the embryonic development up to 2 or 3 days of development (33% of the database) or up to 5 or 6 days of development (67% of the database). The output video score was then fed as input alongside clinical features to a gradient boosting algorithm that generated a second score corresponding to the hybrid model. AUC was computed across 7-fold of the validation dataset for both models. These predictions were compared to those of 13 senior embryologists made on the test dataset.

MAIN RESULTS AND THE ROLE OF CHANCE: The average AUC of the hybrid model across all 7-fold was significantly higher than that of the video model (0.727 versus 0.684, respectively, $P=0.015$; Wilcoxon test). A SHapley Additive exPlanations (SHAP) analysis of the hybrid model showed that the six first most important features to predict pregnancy were morphokinetics of the embryo (video score), oocyte age, total gonadotrophin dose intake, number of embryos generated, number of oocytes retrieved, and endometrium thickness. The hybrid model was shown to be superior to embryologists with respect to different metrics, including the balanced accuracy ($P \leq 0.003$; Wilcoxon test). The likelihood of pregnancy was linearly linked to the hybrid score, with increasing odds ratio (maximum P -value = 0.001), demonstrating the ranking capacity of the model. Training individual hybrid models did not improve predictive performance. A clinic hold-out experiment was conducted and resulted in AUCs ranging between 0.63 and 0.73. Performance of the hybrid model did not vary between TLS or between subgroups of embryos transferred at different days of embryonic development. The hybrid model did fare better for patients older than 35 years ($P < 0.001$; Mann–Whitney test), and for fresh transfers ($P < 0.001$; Mann–Whitney test).

LIMITATIONS, REASONS FOR CAUTION: Participant centers were located in two countries, thus limiting the generalization of our conclusion to wider subpopulations of patients. Not all clinical features were available for all embryos, thus limiting the performances of the hybrid model in some instances.

WIDER IMPLICATIONS OF THE FINDINGS: Our study suggests that considering clinical data improves pregnancy predictive performances and that there is no need to retrain algorithms at the clinic level unless they follow strikingly different practices. This study characterizes a versatile AI algorithm with similar performance on different time-lapse microscopes and on embryos transferred at different development stages. It can also help with patients of different ages and protocols used but with varying performances, presumably because the task of predicting fetal heartbeat becomes more or less hard depending on the clinical context. This AI model can be made widely available and can help embryologists in a wide range of clinical scenarios to standardize their practices.

STUDY FUNDING/COMPETING INTEREST(S): Funding for the study was provided by ImVitro with grant funding received in part from BPIFrance (Bourse French Tech Emergence (DOS0106572/00), Paris Innovation Amorçage (DOS0132841/00), and Aide au Développement DeepTech (DOS0152872/00)). A.B.-C. is a co-owner of, and holds stocks in, ImVitro SAS. A.B.-C. and F.D.M. hold a patent for 'Devices and processes for machine learning prediction of in vitro fertilization' (EP20305914.2). A.D., N.D., M.M.F., and F.D.M. are or have been employees of ImVitro and have been granted stock options. X.P.-V. has been paid as a consultant to ImVitro and has been granted stocks options of ImVitro. L.C.-D. and C.G.-S. have undertaken paid consultancy for ImVitro SAS. The remaining authors have no conflicts to declare.

TRIAL REGISTRATION NUMBER: N/A.

Key words: deep learning / time-lapse incubator systems / embryo evaluation / artificial intelligence / pregnancy / gonadotrophin / machine learning

Introduction

Embryologists routinely use morphokinetic criteria to identify the best embryo to transfer (Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology, 2011). The criteria are based on both morphological traits and the timing of key biological events that can routinely be recorded with time-lapse incubator systems (TLS) for a maximum of 7 days of culture. This routine embryo evaluation can be time-consuming (Veiga *et al.*, 2022) and suffer from inter and intra-operator variability (Adolfsson and Andershed, 2018). In addition, while there is an obvious correlation between embryo quality (as defined by embryologists) and pregnancy success rates, numerous 'good' quality embryos will not lead to a pregnancy, and inversely, 'poor' quality embryos may lead to a clinical pregnancy and live births (Oron *et al.*, 2014). On the other hand, pre-implantation genetic screening for aneuploidy (PGT-A) is considered a diagnostic tool of interest to deselect aneuploid embryos, which might lead to implantation failures and miscarriages (Sanders *et al.*, 2021). However, PGT-A is a costly, time consuming, and invasive procedure and it remains unclear if it benefits equally all patients (Greco *et al.*, 2020), especially subgroups of patients who might have less marked aneuploidy rates such as patients <35 years of age (Demko *et al.*, 2016).

To assist embryologists in evaluating embryos, a growing number of publications have suggested using artificial intelligence (AI), proposing

different approaches (Zaninovic and Rosenwaks, 2020). Some algorithms focus solely on automating the morphological evaluation (Chen *et al.*, 2019; Khosravi *et al.*, 2019; Kragh *et al.*, 2019) or the extraction of morphokinetics events (Dirvanauskas *et al.*, 2019; Feyeux *et al.*, 2020), thus saving time and decreasing inter-operator variability, but leaving embryologists to estimate the embryo quality according to common standards. Other studies use AI to predict the chances an embryo has to develop to a blastocyst (Wong *et al.*, 2010; Liao *et al.*, 2021). Another approach consists of feeding the timing of key biological events to machine learning algorithms, in the hopes that they can use this information to predict the chances of pregnancy or live birth (Zabari *et al.*, 2022) better than embryologists. In contrast, a more objective approach consists of training models that either analyze a single image (VerMilyea *et al.*, 2020) or the entire embryonic development (Tran *et al.*, 2019; Berntsen *et al.*, 2022; Lassen *et al.*, 2022) using pregnancy outcome as a label. This leaves more room for the algorithms to focus on any part of the embryonic development, without biasing them into analyzing solely the information that embryologists focus on (Coticchio *et al.*, 2022). However, analyzing only one final image can only help embryologists that choose to transfer embryos at the blastocyst stage, and seemingly ignores the important and time-consuming task of analyzing the kinetics of the embryo (Meseguer *et al.*, 2011; Veiga *et al.*, 2022).

While additional groups have described deep learning algorithms trained on videos of developing embryos (Sawada *et al.*, 2021; Yang

et al., 2022), none seem to have combined them with multiple variables describing the characteristics of the patient or the cycle. This is necessary to adjust the chances an embryo has to lead to a pregnancy, in the context, for example, of endometrial receptivity as it is known the interactions between the embryo and the endometrium play an important role in embryo implantation (Lessey and Young, 2019). Erlich et al. (2021) only used the age of the egg donor or mother, while Enatsu et al. (2022) analyzed 12 clinical features (including age, anti-Müllerian hormone (AMH), endometrial thickness) but in combination with single images of the blastocyst. The aim of this study was to test whether analyzing a large subset of clinical features improved the performances of a deep learning algorithm that predicts the likelihood of pregnancy of an embryo based on its kinetics. To the best of our knowledge, this is the first algorithm that personalizes the evaluation of embryo kinetics with a set of 31 clinical features, across different time-lapse incubators and on embryos transferred at different stages of development. Our aim is to provide IVF practitioners with a complementary automated solution that helps them predict which embryo can lead to a clinical pregnancy.

Materials and methods

Data collection

This study was performed on data from 14 clinics in France and Spain, corresponding to IVF cycles completed between the year 2016 and 2022.

The data, consisting of videos of embryos and 31 clinical variables describing the patients (or donors when relevant) and their IVF treatment, was collected after removing any direct identifiers of the patient. This medical information consisted of clinical variables such as age, BMI, hormonal levels, and type of stimulation, which are all detailed in Supplementary Table S1. The videos were recorded using one of the following TLS: Embryoscope[®] or Embryoscope+[®] (Vitrolife, Västra Frölunda, 421000, Sweden), GERI[®] (Genea BiomedX, Sydney NSW 2000, Australia), or MIRI[®] (ESCO Medical, Egå, 8250 Denmark).

Data from a total of 55 077 embryos was collected, but only 9986 embryos (Table I) were used to train the algorithms (including 9537 embryos with known pregnancy data) and 447 for testing (all having known pregnancy data, Table II). Only 192 embryos used in training (<2% of known pregnancy data) had been biopsied for PGT-A testing, and <3% corresponded to oocyte donations.

Each embryo and thus, video, was linked to: its transfer decision (discarded, frozen, transferred fresh, or transferred frozen); its clinical pregnancy outcome; and the clinical features of the cycle related to the patient (or donor when relevant, Supplementary Table S1). Clinical pregnancies were categorized into a positive (FH+) or a negative (FH-) clinical pregnancy, using the detection of a fetal heartbeat (FH) via ultrasound (6–8 weeks after transfer) as a measure of clinical pregnancy.

Retrospective clinical trial

To compare the AI model performances to that of experts, 13 senior embryologists from nine different IVF centers were asked to analyze retrospective data through an annotation platform available online. Each one analyzed a subset of ~130 videos randomly selected from the 447 videos present in the test set (Table II and Supplementary

Table I Description of the complete training and validation database.

Training and validation database (9986 human embryos)	
Transfer decision (number of embryos (%)) [†]	
Discarded with extremely poor quality	449 (4%)
Transfers with known FH outcome	9537 (96%)
Time-lapse system	
MIRI [®]	39%
GERI [®]	32%
EMBR–EMBR+ [®]	29%
IVF technique	
Conventional IVF	23%
ICSI	63%
Not reported	14%
Transfer day	
Day 2	15%
Day 3	18%
Day 4	19%
Day 5	39%
Day 6	9%
Type of transfer	
Fresh	58%
Frozen	34%
Not reported	8%
Main clinical features (total 31)	
Oocyte age [years], mean ± SD (min–max)	35 ± 5 (19–54)
Woman BMI [kg/m ²], mean ± SD (min–max)	23.4 ± 4.3 (14–45)
AMH [ng/ml], mean ± SD (min–max)	2.95 ± 2.62 (0.01–29)
FSH [IU/l], mean ± SD (min–max)	7.29 ± 3.13 (0.4–67.0)

*Extremely poor embryos are described in 'Data preparation'.

[†]Transfers with known outcome correspond to single transfers or double transfers where the outcome applies to both embryos (i.e. two FH+ or no FH+ at all).

FH: fetal heartbeat, AMH: anti-Müllerian hormone.

Fig. S1). Some videos were seen by more than one expert. Each embryologist's performance was compared to that of the model based only on the subset of embryos they had seen.

Embryologists were asked to predict the clinical pregnancy outcome of the embryo as positive or negative, with and without the contextualizing clinical data. They also had to grade the embryo quality: poor, fair, or good, according to morphokinetic criteria (Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology, 2011). Note that the embryo quality grading was voluntarily generic so as to encompass practices from different countries. Future studies will benchmark the algorithm with other grading systems (e.g. Gardner, ASEBIR (Arday et al., 2008; Gardner and Schoolcraft, 1999), or with respect to morphological traits at the blastulation stage). The test set was not a random sampling of the embryos with known pregnancy data as it needed to contain a stable amount of data coming from each

Table II Test set description.

Test set (447 human embryos)	
Time-lapse system	
EMBR-EMBR ⁺	36%
GERI [®]	32%
MIRI [®]	32%
IVF method*	
Conventional IVF	26%
ICSI	72%
Unknown	2%
Frozen transfers	
45%	
Transfer day	
Day 2	20%
Day 3	24%
Day 5	37%
Day 6	20%
Embryo quality	
Poor	23%
Fair	41%
Good	36%
Main clinical features	
Oocyte age [years], mean ± SD (min–max)	35 ± 5 (20–43)
BMI woman [kg/m ²], mean ± SD (min–max)	23.2 ± 4.5 (15–42)
AMH2 [ng/ml], mean ± SD (min–max)	3.04 ± 3.08 (0.33–22)
FSH3 [IU/l], mean ± SD (min–max)	7.1 ± 1.9 (1–16.7)

*IVF method is not always reported in the data explaining the presence of unknown data.

AMH: anti-Müllerian hormone.

TLS (Table II). No embryo coming from the same egg retrieval could be found both in the test set and the training or validation set.

Data preparation

Given that data came from different clinics, countries, and TLS, a normalization pipeline was developed to homogenize both the videos and the clinical data. The goal of this pipeline was primarily to facilitate the learning process and to avoid biases. The normalization pipeline of videos consisted of the following steps: temporal normalization, embryo cropping, resizing, removal of the final empty frames as well as luminosity correction.

Videos that covered <24 h of embryonic development were removed (~5% of the training and validation database). All videos were uniformly sub-sampled to the maximum acquisition rate (i.e. 20 min). Each frame of the video was cropped over time to a 256 × 256 image centered around the embryo using YOLO (Redmon et al., 2015) to minimize exposure to artifactual information outside of the embryo. The YOLO model was evaluated on a validation set of 300 frames and reached a precision and recall of 0.961 and 0.992, respectively.

While the vast majority of the validation and training data corresponded to known pregnancy data, an exception was made for a small portion of discarded embryos curated by a senior embryologist: these extremely poor quality embryos were identified as never being able to lead to a pregnancy (e.g. immature eggs, stopped in their development at any cleaved stage; Supplementary Fig. S2). They only represented 4.5% of the validation and training dataset, and were never taken into account in the performances reported in this paper, other than in the discussion to estimate how easy it would be to automatically identify these embryos. Indeed, the focus of this study was to train a model only on transferred embryos, to not reproduce the potential bias of the embryologist. The risk would be that the AI always discards poor embryos that could have led to a pregnancy (Oron et al., 2014; Kirillova et al., 2020). Therefore, no assumptions were made about the outcome of the vast majority of discarded embryos.

Training of the deep learning and machine learning models

To predict the pregnancy outcome of the embryo, a deep learning model (3D ConvNet network with a ResNet backbone (He et al., 2015)) was built using PyTorch (Paszke et al., 2019). This architecture was proven to be very effective for video-classification tasks (Tran et al., 2014). Using a depth of 50 layers, the model is made of 27.2 M parameters that are optimized through a training loop. The final output is a score ranging between 0 and 1 and is evaluated using a binary cross entropy loss. Prior to training, videos are resized to 128 × 128 to speed up the learning.

To deal with the limited GPU memory, 64 frames are uniformly sampled across the video using a stride of three frames. The developmental time of the embryo between two frames corresponded to 2 h. If the video spanned <128 h, it was padded with black images up to the 64th frame. Gaussian blur was applied during training, as well as Gaussian noise and color jittering (contrast, brightness, saturation). Finally, the images were rotated at a random angle, cropped randomly to 87% of the field of view and flipped horizontally.

The training and validation dataset was split into seven cross-validation folds to ensure a ratio of ~85%/15% between training and validation. No embryo from the same egg retrieval could be found both in the validation and the training subset. Each training fold was used to train a model and metrics were computed on each validation set. The model was trained during 86 722 mini-batches, corresponding to 50 epochs, using a batch size of 9 on 4 T4 GPUs. The initial learning rate was set at 0.05 and is decreased by 1e-1 first at Epoch 5 and a second time at Epoch 15. This schedule and base learning rate were found to yield the most stable learning curve and the highest validation performance. To control overfitting, a dropout rate of 0.5 was used. This hyper-parameter controls the ratio of neurons in the network that will be ignored during training and is used to reduce overfitting. Transfer learning was used as a warm-start, using a set of weights originally pre-trained on the dataset Kinetics (Kay et al., 2017). The final fully connected layer was replaced to match the number of classes in our dataset (N = 1).

Clinical features were pre-processed and fed along with the video score provided by the 3D ConvNet model to a gradient boosted decision tree algorithm (XGBoost), designated as the hybrid model. Hyperparameters were automatically adjusted in order to maximize

F1-score from the validation sets. The hybrid model was trained on the same 7-fold stratified cross-validation, and provided a score, ranging from 0 to 1. All performances reported on the test set correspond to models (hybrid or video) retrained on the entire training and validation database to leverage all the data available.

Hybrid models were trained on the same seven splits described before, either on each clinics' data (i.e. customized) or on data from all clinics (i.e. generic). Performances were compared between the customized and generic model on data for each clinic.

To identify the importance of clinical features on the final score, an SHAP explainer model was fitted on the test set, using all the available clinical features, including the video score obtained thanks to the 3D convolutional neural network (Lundberg et al., 2020). This value is computed using a framework based on co-operative game theory and assesses the impact of each feature in context with every other feature available. SHAP values were computed for each feature and averaged across all embryos from the test set.

A clinic hold-out validation study was carried out to investigate the generalizability of the hybrid model as suggested by Kragh and Karstoft (2021), similarly to Bermtsen et al. (2022). The hybrid model was trained on all clinics but one, and then evaluated on the data from the clinic that was left out. The top 10 features according to the SHAP analysis were computed for each fold, alongside the average percentage of missing values for those features of the evaluated clinic. Clinics with less than 250 embryos with known pregnancy data and <50% of data within the top 10 clinical features were excluded from this analysis.

Model performances

The performance of the models was first assessed using the receiver operating characteristic (ROC) curve generated by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across all possible thresholding values between 0 and 1. The higher the AUC, the more favorable the trade-off between sensitivity and specificity. Balanced accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated by selecting a specific threshold, following the F1 score maximization as a measure of the best compromise between sensitivity (recall) and PPV (precision). The Matthew correlation coefficient (MCC) was also computed, which is a measure of the differences between the predicted and the actual values. A value of 1 indicates perfect agreement, while a value of 0 indicates a random prediction. The optimal threshold for F1-score was found on the concatenation of all validation sets. Unless specified, performances were averaged across all validation sets. They were reported on the test set only when comparing the algorithm to embryologists in the retrospective trial.

The video score was binned in four separate categories to test whether there was a correlation between scores and pregnancy rates, and whether different groups of embryos could be distinguished as a function of their likelihood to lead to a pregnancy. Negative predictions were binned into *Not Recommended* (0–0.14) and *Not Favorable* (0.15–0.48) categories, and positive predictions into *Favorable* (0.49–0.68) and *Recommended* (0.69–1).

Statistical analysis

The Wilcoxon test was used to assess statistically significant differences for quantitative data. A Spearman's rank correlation was calculated to assess

the relation between the video score and the grade of the embryo. A logistic regression was fitted between the pregnancy outcome and the bins of hybrid score and was used to compute the odds ratios (OR). Mann-Whitney *U* test was used to compare AUCs between different groups. A Bonferroni correction was used for comparisons across more than two groups. All statistical comparisons were two sided and a significance threshold of 5% for the *P*-value was used. Averages are shown either with their SD or their interval of confidence (95% CI). The Python package SciPy (v.1.8.1) was used to conduct all statistical tests (Virtanen et al., 2020).

Results

The video score correlates with embryo quality

Figure 1 shows the distribution of video scores for each embryo's grade. Out of the 447 embryos from the test set, $N = 102$ were labeled as poor, $N = 184$ as fair quality, and $N = 161$ as good quality by the embryologist's majority vote (see Materials and methods). Average scores for poor quality embryos tended to be lower (0.23 ± 0.20), whereas good embryos corresponded to much higher scores (0.50 ± 0.22). The correlation between the video score and embryo quality was statistically significant (Spearman's test, $\rho = 0.45$, *P*-value < 0.001).

Analyzing clinical features in addition to embryo development improves performances

The results of the 3D-ResNet model trained only with videos were compared to the hybrid model that analyzed both the output of the 3D-ResNet and the clinical features (Fig. 2A). The hybrid model outperformed the 3D ResNet model, with a statistically significant relative increase of 6.27% in AUC ($P = 0.0011$; Wilcoxon test). Analyzing both the videos and the clinical features yielded an average AUC of 0.727 ± 0.012 versus 0.684 ± 0.016 when only visual information was taken into account (Fig. 2B). Performances measured on each fold can be found in Supplementary Table SII.

Figure 3A shows the relative contribution of the top 10 features to the final hybrid score, as identified by XGBoost. The larger the SHAP value is, the larger the impact a given feature on the final score. The results showed that the video score, which is the output score of the 3D-ResNet and reflects the morphokinetics of the embryo, was the most important feature, with a mean absolute contribution of 0.57 ± 0.40 . Oocyte age was the second most important feature with a mean absolute SHAP value of 0.32 ± 0.36 . The 22 clinical features present in our database (Supplementary Table SI) and not shown in Fig. 3 accounted for 0.42 ± 0.01 of the SHAP contribution, with none of the individual features displaying a value higher than 0.037.

The hybrid score correlates with pregnancy rate

In order to test whether the scores of the hybrid model increased with the likelihood of pregnancy, they were binned into four categories. Figure 3B shows the ratio of successful pregnancy for each group of scores. The increasing ORs between the first category and the rest

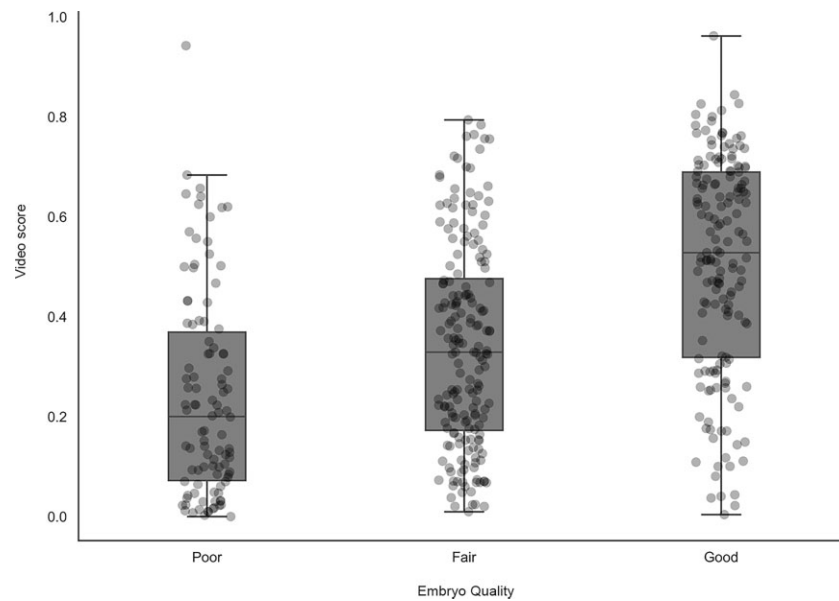


Figure 1. Boxplots of the video scores according to embryo grade computed on the test set. Each dot corresponds to one human embryo. Out of the 447 embryos from the test set, $n = 102$ were labeled as poor, $n = 184$ as fair quality, and $n = 161$ as good quality by the embryologist's majority vote.

indicates a linear relationship between the score and the likelihood of pregnancy. This suggests the ability of the model to rank embryos to a certain degree according to their potential to lead to a pregnancy. It is important to always keep in mind the imbalance of the dataset as a measure of what a random prediction would result in. In this case, a random model would result in a 27% pregnancy rate across all bins.

The hybrid algorithm generalizes to different clinical contexts but fares better in cases of fresh transfers and in women aged 35 years or more

There was no statistical difference between time-lapse microscopes ($P > 0.046$, Bonferroni correction), or in terms of day of transfer ($P = 0.71$; Table III). Performances across different clinical scenarios were computed (Table IV), showing a statistically significant difference (P -value < 0.001) between the fresh and frozen transfer groups in terms of AUC (0.76 ± 0.02 versus 0.67 ± 0.01 , respectively). Performances were higher (P -value < 0.001) for women aged 35 years or more (AUC 0.74 ± 0.02) compared to women younger than 35 (AUC 0.68 ± 0.01).

The resulting AUCs for each clinic of the hold-out clinic experiments, with the 95% CI, are shown in Table V. Only centers that had at least 50% of the top 10 features associated with their videos are shown.

Supplementary Table SIII shows that personalizing the hybrid model to each clinic's data did not significantly improve performances (P -value > 0.303 ; DeLong test) for any of the clinics participating in this study,

suggesting clinics would not benefit from an algorithm trained specifically on their data.

Performances were also evaluated against a diverse panel of embryologists ($N = 13$) on the test set based on the threshold optimized on the validation set ($t = 0.49$).

The performances obtained showed that the hybrid model was significantly better than the embryologists in terms of specificity (0.73 ± 0.05 versus 0.56 ± 0.10 , P -value < 0.001) and PPV (0.40 ± 0.08 versus 0.34 ± 0.07 , P -value = 0.001), but also significantly lower in terms of sensitivity (0.50 ± 0.06 versus 0.59 ± 0.07 , P -value = 0.008). NPV of the hybrid model was slightly above that of the embryologist (0.80 ± 0.02 versus 0.79 ± 0.03 , P -value = 0.63), with no statistical difference found. Using the MCC, the hybrid model was again significantly higher than the embryologists (0.21 ± 0.08 versus 0.14 ± 0.10 , P -value = 0.021), and also with respect to the balanced accuracy (0.61 ± 0.04 versus 0.58 ± 0.05 , P -value = 0.03), two metrics known to be independent of the imbalance of the predicted class (Fig. 4). Individual performances per embryologist are shown in Supplementary Table SIV.

Discussion

To the best of our knowledge, this is the first study to predict the likelihood of fetal heartbeat pregnancy of embryos recorded with different TLS, transferred at different times in the development of the embryos, and using a large panel of clinical features that also describe the patient and their IVF treatment. This study paves the way for a more holistic and personalized approach to predicting fetal heart rate pregnancy.

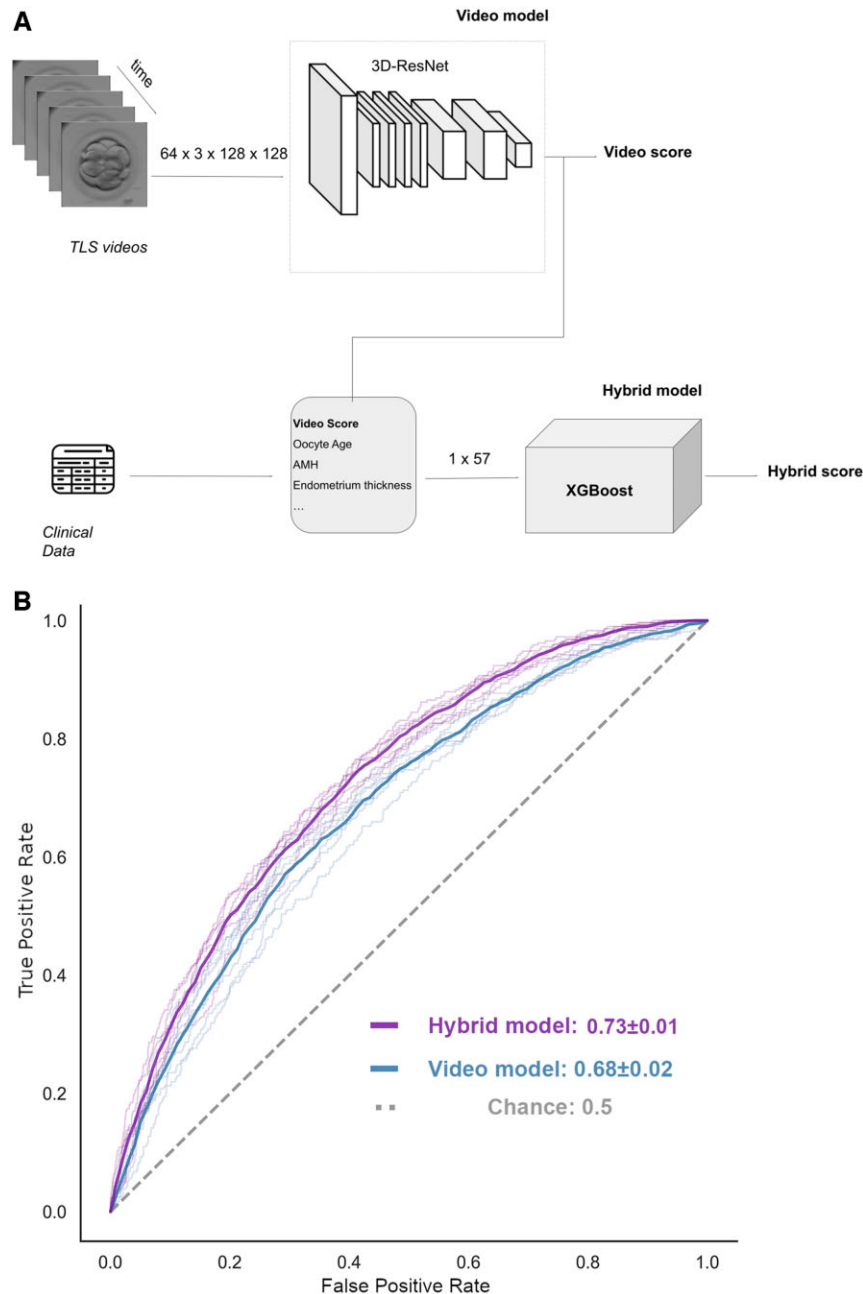


Figure 2. Results of the 3D-ResNet model trained only with videos were compared to the hybrid model that analyzed both the output of the 3D-ResNet and the clinical features. (A) Architecture of the two models. (B) AUC of the video model and hybrid model averaged on the 7-fold. TLS: time-lapse system, AMH: anti-Müllerian hormone, XGBoost: a gradient boosted decision tree algorithm designated as the hybrid model.

The results of this study show that taking clinical parameters into account significantly increased the AUC by 6% (Fig. 2B), similarly to Enatsu et al. (2022) who showed an increase of 4.5%. Their study was, however, applied to static images and was limited to the use of 12 clinical features, in contrast with the 31 types of variables present in the database of this study. Interestingly, their SHAP analysis also

identified the age of the oocyte as the second most important feature after the blastocyst image, equivalent to the video score in this study. They also show AMH as another key predictive feature of pregnancy outcome, which was not ranked as high in our hands. AMH has been suggested as an indicator of ovarian reserve (Iwase et al., 2016); perhaps it was not prioritized as much by this model because the number

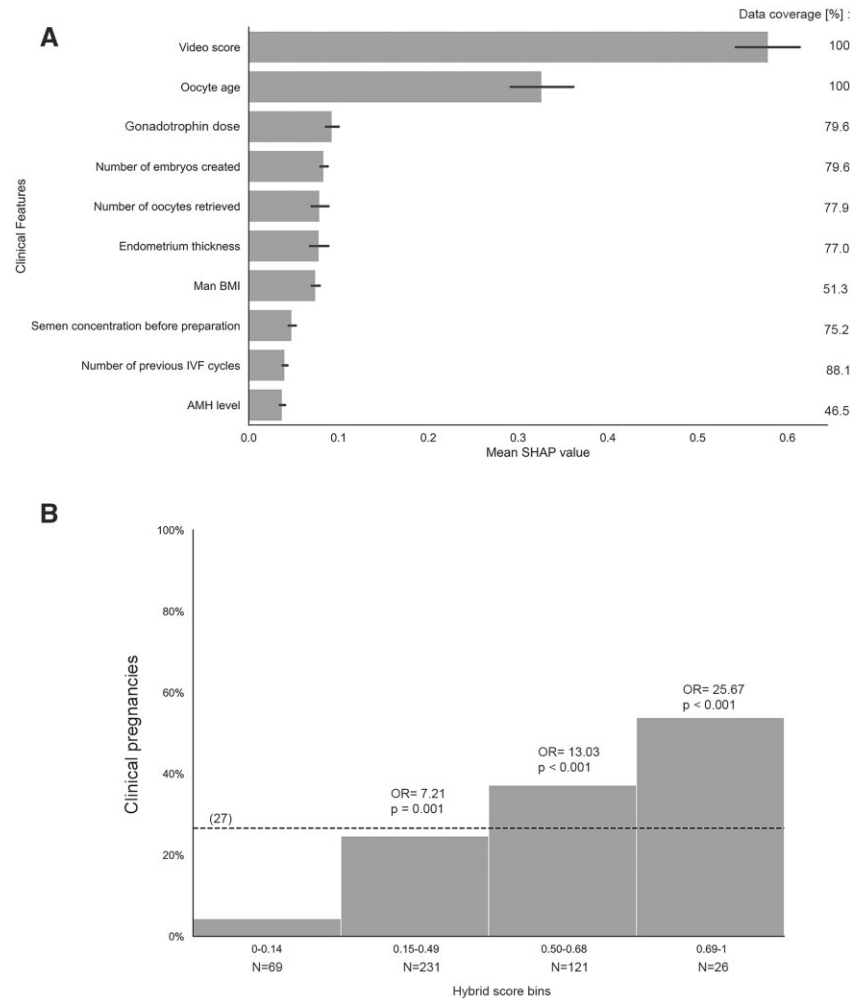


Figure 3. Relative contribution of the top 10 features to the final hybrid score, as identified by a gradient boosted decision tree algorithm, XGBoost. (A) The percentage of videos that had each of the corresponding clinical features is shown. Ranking of the top 10 features by mean SHAP values computed on the test set, i.e. their synergistic contribution to the final score. The larger the SHAP value is, the larger the impact a given feature on the final score. **(B)** Percentage of successful clinical pregnancy for each range of hybrid scores on the test set. The dashed horizontal line indicates the prevalence of the fetal heartbeat (FH) label in the dataset. A logistic regression was fitted between the variable FH and each category of scores. The corresponding odds ratios with their *P*-values are shown for each category compared to the first bin. SHAP: SHapley Additive exPlanations, an explainer model, OR: odds ratio, AMH: anti-Müllerian hormone.

of oocytes retrieved was also available and used as another measure of the ovarian reserve (Yih *et al.*, 2005). The percentage of videos that had each of the corresponding clinical features is shown in Fig. 3A. Results suggest that a feature's importance is not only correlated with its presence in the test set. Based on the top 10 predictive biomarkers of this study, it seems as if the model prioritized a combination of features concerning: sperm (male BMI and semen concentration); oocyte quality (oocyte age, AMH, and gonadotrophin dosage); and the patient's receptivity (endometrium thickness), which have previously been reported in some studies to be linked to IVF outcomes (Stadtmauer *et al.*, 1994; Gleicher *et al.*, 2016; Mushtaq *et al.*, 2018; Cimadomo *et al.*, 2018; Craciunas *et al.*, 2019). Future experiments are needed to discriminate between different types and starting doses

of gonadotrophins, which could affect clinical pregnancy outcome in different ways. Interestingly, the length of gonadotrophin exposure seemed to have minimal impact on the pregnancy outcome prediction (SHAP value = 0.0078 ± 0.0114), in agreement with other studies (Martin *et al.*, 2006).

While it is not possible to directly compare performances with other publications that have used a different dataset to evaluate their performances, the AUCs of both video and hybrid models from this study are comparable to that of other groups that have trained their algorithms to predict the fetal heartbeat. For example, Lassen *et al.* (2022) and Erlich *et al.* (2022) have reported AUCs ranging from 0.621 to 0.708 on Day 5. The original study of Tran *et al.* (2019) reported an extremely high AUC of 0.93, but only because they

Table III Performances of the video model in different subgroups.

Group	Number of known pregnancy data	AUC	95% CI
Transfer day			
Day 2–Day 3	3166	0.66	0.61–0.71
Day 5–Day 6	6371	0.65	0.64–0.67
Time-lapse system			
MIRI [®]	3369	0.69	0.66–0.72
GERI [®]	2738	0.65	0.64–0.67
EMBR/EMBR+ [®]	3130	0.7	0.65–0.74

Values are computed over seven cross-validation splits.

Table IV Performances of the hybrid model across different clinical scenarios.

Group	Number of known pregnancy data	AUC	95% CI
Transfer type			
Frozen	3217	0.67*	0.66–0.68
Fresh	6320	0.76	0.74–0.78
Oocyte age [years]			
Women < 35 years	3824	0.68*	0.66–0.69
Women ≥ 35 years	4619	0.74	0.72–0.76
Fertilization method			
Conventional IVF	2112	0.72	0.69–0.76
ICSI	5977	0.73	0.71–0.74

The asterisks show statistical differences between subgroups (Mann–Whitney U test). Values are computed over seven cross-validation splits.

Table V Clinic hold-out AUC results and 95% CI.

Clinic [*]	Number of KID embryos	% Missing top 10 clinical features ^{**}	% FH+	AUC Hybrid Model	95% CI
1	326	38%	29%	0.63	0.57–0.69
2	1247	32%	16%	0.7	0.67–0.73
3	1671	12%	25%	0.68	0.66–0.71
4	2981	9%	19%	0.72	0.7–0.73
5	620	7%	23%	0.7	0.65–0.74
6	694	1%	17%	0.69	0.65–0.72

KID: known implantation. FH+: positive fetal heartbeat.

Each row shows data pertaining to the clinic that was left out from training.

*ID of the clinic excluded from training.

**Data available from the 10 most important clinical features described in Fig. 3.

trained and validated their algorithms on a disproportionate number of discarded embryos; this has been acknowledged as being an easier task that facilitates higher AUCs (Chavez-Badiola et al., 2020; Kan-Tor

et al., 2020; Tran et al., 2019, 2020a, 2020b). When the same group looked only at known implantation data, the AUC dropped from 0.94 to 0.708 (Lassen et al., 2022). We have therefore trained and evaluated separately our algorithms on a small subset of discarded embryos that were unquestionably never leading to a pregnancy and also obtained extremely high performances to recognize them, with an accuracy of 94% using the hybrid model. This can become important to one day use AI to automate the entire process of embryo evaluation.

It is, however, possible to benchmark the performance of the hybrid model with embryologists by comparing their predictions on a common dataset. The results demonstrate a clinical superiority based here on a higher specificity (Fig. 4), suggesting the model is more often right in de-selecting embryos that cannot lead to a pregnancy. Here the threshold was chosen following the F1-score maximization; the threshold could have been optimized differently to prioritize other metrics. Choosing a threshold is all about striking a compromise, and it could therefore be tailored at the clinic-level if a center prefers to optimize the ability to better select embryos that can lead to a pregnancy, at the risk of having a lower specificity. Ongoing studies are evaluating the abilities of this same model to better select ‘poor’ embryos that could lead to a pregnancy, which could be an interesting application for embryologists who might be less used to transferring them.

Much like these existing algorithms, our algorithms are inherently biased because they were not trained on the many embryos that were never transferred, because they lack pregnancy data. However, the database was longitudinal enough that 46% of the embryos used in training correspond to embryos transferred after a first failed transfer. This ensures that the algorithm has not only seen the embryos that embryologists tend to transfer first, and thereby engrains some crucial diversity in the type of embryo the model is used to evaluate.

Thanks to the diversity of the database, it was possible to demonstrate the versatility of the described algorithm. No significant difference in performances across time-lapse microscopes was observed. The clinic hold-out study shown in Table V demonstrated that, for a clinic that was never seen before during training, the AUC of the hybrid model could range from 0.63 to 0.72. Lower performances were observed for clinics that did not have as much associated clinical data. Notably, the only Spanish center eligible for this hold-out experiment showed the lowest AUC. This probably is because practices are known to differ between France and Spain (e.g. more or less oocyte donors or PGT-A testing). Therefore, excluding even more Spanish data from training (which represented 11% of the validation and training database) could explain the lower performance. This emphasizes the need to maintain a diverse database representative of different practices over time. In addition, results in Supplementary Table SIV show that none of the participating centers would have benefited from having a hybrid model trained solely on their data. It was also observed that the predictive capabilities of the video model were not significantly lower for embryos transferred earlier in embryonic development, suggesting this algorithm can also benefit clinics that choose to transfer embryos earlier in development. This suggests that these algorithms might be identifying key early events, in line with other publications (Coticchio et al., 2022), which have highlighted early embryonic events as being predictive of implantation rate. Ongoing work from our group is looking into explainability models that could help us better understand what events are driving the deep learning algorithm. Just like Berntsen et al. (2022) and Lessen et al. (2022), this

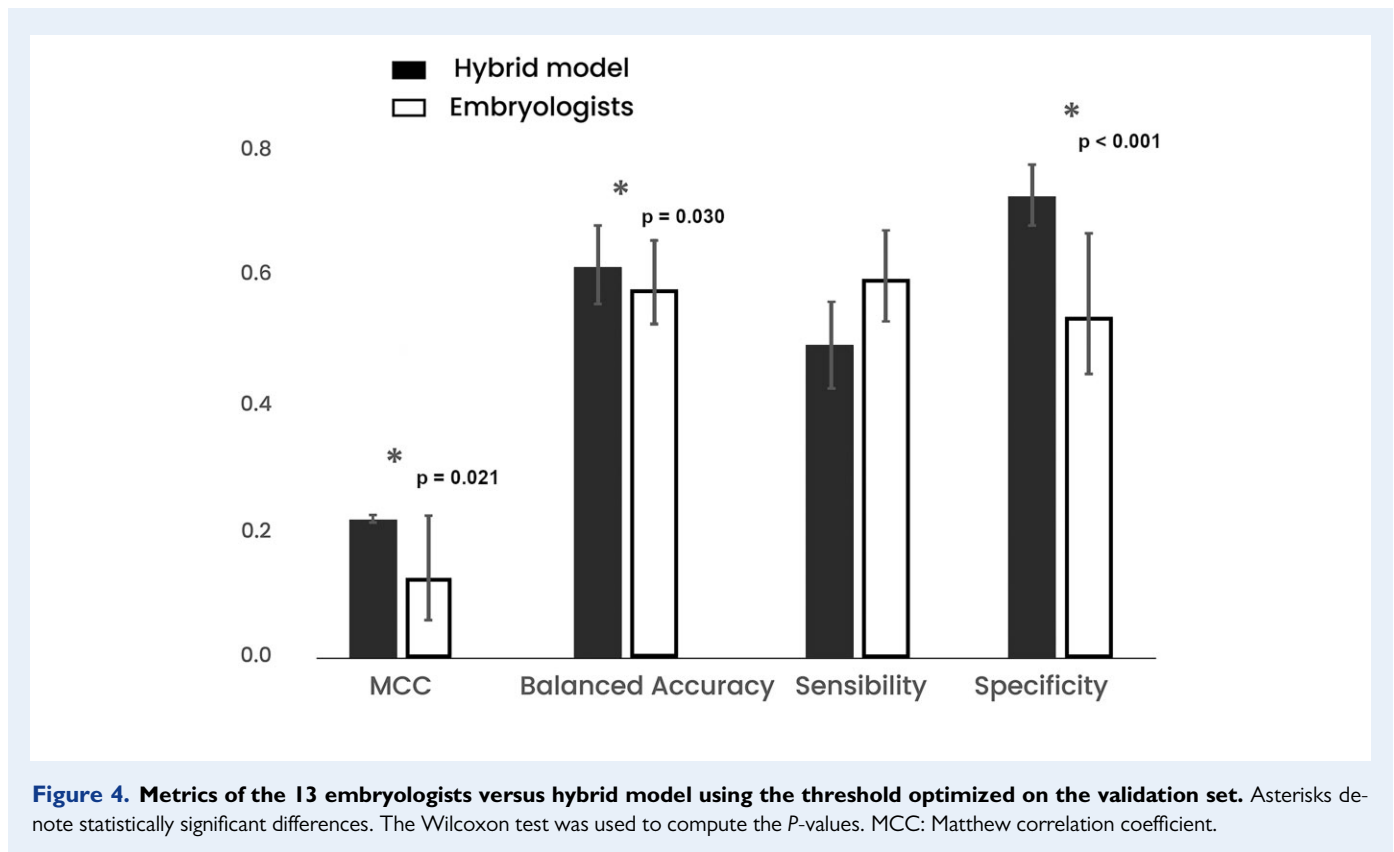


Figure 4. Metrics of the 13 embryologists versus hybrid model using the threshold optimized on the validation set. Asterisks denote statistically significant differences. The Wilcoxon test was used to compute the *P*-values. MCC: Matthew correlation coefficient.

model can better predict the likelihood of pregnancy for fresh transfers. This might have been due to the fact that the clinical features used were measured at the time of egg retrieval, without taking into account some variables that might have changed at the time of frozen transfers (e.g. endometrium thickness). However, training a model only on data from frozen embryos did not improve performances significantly either. It is to be noted that a very small portion of the data corresponded to oocyte donations; in those cases, the age of the donor was taken into account to reflect oocyte and thus embryo quality. Future studies will combine both clinical variables from the donor and the patient to also account for their impact on the patient's receptivity. There was a decrease in AUC for younger patients, as found by [Erlich et al. \(2022\)](#). They hypothesized that it could be due to this sub-group having more infertility due to non-embryonic causes, thus resulting in noisy labels where a high potential embryo does not lead as often to a successful pregnancy. Noisy labels are often an issue for a deep learning model as it can damage its final accuracy. There are machine learning techniques ([Song et al., 2022](#)) that can mitigate the effect of such noisy labels, which will also be the focus of future investigations.

The task of embryologists is first and foremost to rank embryos and to transfer the one(s) deemed to have the highest chance of leading to a pregnancy. Results in [Fig. 3](#) show that in fact the hybrid score has ranking abilities that can help the embryologists bin embryos into different likelihoods of pregnancy. Future studies will have to evaluate the ranking abilities in the context of cohorts of embryos to understand their impact on the cumulative pregnancy rate and time to pregnancy ([Diakiw et al., 2022](#)). In addition, the described hybrid score

adjusts the chances an embryo has to lead to a pregnancy in context with the patient and treatment characteristics, which can be difficult for embryologists as it encompasses many variables. This can help with better treatment expectations and management, including better timing of transfer, especially if variables concerning the endometrium at transfer are evaluated. This could also help identify the clinical drivers of this success, which can be helpful in case any of them are actionable for future transfers or other patients with similar characteristics. At any rate, additional types of clinical features will be collected to keep improving predictive performances, while acknowledging that a ceiling in performances will be reached given that the likelihood of pregnancy does depend to some extent on factors that are currently not captured by any available data (e.g. impact of the embryo transfer procedure).

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Data availability

The embryo videos and other patient data collected in this study are not publicly available owing to reasonable ethics and privacy concerns, and are not redistributable. For any interested collaborators, please contact the corresponding author. The AI model developed in this article is available for commercial use as part of ImVitro's software. The computer code developed is not publicly available owing to commercial restrictions.

Acknowledgements

The authors want to thank Dr Emmanuel Chamorey for his help with our statistical analysis. Dr Jessica Vandame for her help labeling embryos.

Authors' roles

A.D., N.D., F.D.M., M.M.F., and A.B.-C. conceived the study, designed methodology. A.B.-C. was responsible for project management and supervision of research activity. A.D. was responsible for data curation, performing the research, formal analysis, and AI development. F.D.M. and N.D. were responsible for clinical data curation and performed research on the impact of clinical features. A.D., N.D., and A.B.-C. wrote the manuscript. X.P.V., D.N., M.F.-B., and L.C.-D. were involved in the design of the methodology. All the authors contributed to the data collection, interpretation, and the review and editing of the final manuscript.

Funding

Funding for the study was provided by ImVitro with grant funding received in part from BPIFrance (Bourse French Tech Emergence (DOS0106572/00), Paris Innovation Amorçage (DOS0132841/00), and Aide au Développement DeepTech (DOS0152872/00)).

Conflict of interest

A.B.-C. is a co-owner of, and holds stocks in, ImVitro SAS. A.B.-C. and F.D.M. hold a patent for 'Devices and processes for machine learning prediction of *in vitro* fertilization' (EP20305914.2). A.D., N.D., M.M.F., and F.D.M. are or have been employees of ImVitro and have been granted stock options. X.P.-V. has been paid as a consultant to ImVitro and has been granted stocks options of ImVitro. L.C.-D. and C.G.-S. have undertaken paid consultancy for ImVitro SAS. The remaining authors have no conflicts to declare.

References

Adolfsson E, Andershed AN. Morphology vs morphokinetics: a retrospective comparison of inter-observer and intra-observer agreement between embryologists on blastocysts with known implantation outcome. *JBRA Assist Reprod* 2018;**22**:228–237.

Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology. The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Hum Reprod* 2011;**26**:1270–1283.

Ardoy M, Caderón G, Cuadros J, Figueroa MJ, Herrero R, Moreno JM, Ortiz A, Prados F, Rodríguez L, Santalo Pedro J et al. Criterios ASEBIR de valoración morfológica de oocitos, embriones tempranos y blastocistos humanos. *Cuadernos de Embriología Clínica* 2008; **11**:1–59.

Berntsen J, Rimestad J, Lassen JT, Tran D, Kragh MF. Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences. *PLoS One* 2022;**17**:e0262661.

Chavez-Badiola A, Mendizabal-Ruiz G, Flores-Saiffe Farias A, Garcia-Sanchez R, Drakeley AJ. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer [Review of Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer]. *Hum Reprod* 2020;**35**:482.

Chen T-J, Zheng W-L, Liu C-H, Huang I, Lai H-H, Liu M. Using deep learning with large dataset of microscope images to develop an automated embryo grading system. *FandR* 2019;**01**:51–56.

Cimadomo D, Fabozzi G, Vaiarelli A, Ubaldi N, Ubaldi FM, Rienzi L. Impact of maternal age on oocyte and embryo competence. *Front Endocrinol (Lausanne)* 2018;**9**:327.

Coticchio G, Borini A, Zacà C, Makrakis E, Sfontouris I. Fertilization signatures as biomarkers of embryo quality. *Hum Reprod* 2022;**37**:1704–1711.

Craciunas L, Gallos I, Chu J, Bourne T, Quenby S, Brosens JJ, Coomarasamy A. Conventional and modern markers of endometrial receptivity: a systematic review and meta-analysis. *Hum Reprod Update* 2019;**25**:202–223.

Demko ZP, Simon AL, McCoy RC, Petrov DA, Rabinowitz M. Effects of maternal age on euploidy rates in a large cohort of embryos analyzed with 24-chromosome single-nucleotide polymorphism-based preimplantation genetic screening. *Fertil Steril* 2016; **105**:1307–1313.

Diakiw SM, Hall JMM, VerMilyea M, Lim AYY, Quangananurug W, Chanchamroen S, Bankowski B, Stones R, Storr A, Miller A et al. An artificial intelligence model correlated with morphological and genetic features of blastocyst quality improves ranking of viable embryos. *Reprod Biomed Online* 2022;**45**:1105–1117.

Dirvanauskas D, Maskeliunas R, Raudonis V, Damasevicius R. Embryo development stage prediction algorithm for automated time lapse incubators. *Comput Methods Programs Biomed* 2019;**177**:161–174.

Enatsu N, Miyatsuka I, An LM, Inubushi M, Enatsu K, Otsuki J, Iwasaki T, Kokeguchi S, Shiotani M. A novel system based on artificial intelligence for predicting blastocyst viability and visualizing the explanation. *Reprod Med Biol* 2022;**21**:e12443.

Erlich I, Ben-Meir A, Har-Vardi I, Grifo JA, Zaritsky A. Solving the "right" problems for effective machine learning driven *in vitro* fertilization. medRxiv 2021.10.07.21264503. <https://doi.org/10.1101/2021.10.07.21264503>.

Erlich I, Ben-Meir A, Har-Vardi I, Grifo J, Wang F, McCaffrey C, McCulloh D, Or Y, Wolf L. Pseudo contrastive labeling for predicting IVF embryo developmental potential. *Sci Rep* 2022;**12**:2488.

Feyeux M, Reignier A, Mocaer M, Lammers J, Meistermann D, Barrière P, Paul-Gilloteaux P, David L, Fréour T. Development of automated annotation software for human embryo morphokinetics. *Hum Reprod* 2020;**35**:557–564.

Gardner DK, Schoolcraft WB. Culture and transfer of human blastocysts. *Curr Opin Obstet Gynecol* 1999;**11**:307–311.

Gleicher N, Kushnir VA, Sen A, Darmon SK, Weghofer A, Wu Y-G, Wang Q, Zhang L, Albertini DF, Barad DH. Definition by FSH, AMH and embryo numbers of good-, intermediate- and poor-prognosis patients suggests previously unknown IVF outcome-determining factor associated with AMH. *J Transl Med* 2016;**14**:172.

- Greco E, Litwicka K, Minasi MG, Cursio E, Greco PF, Barillari P. Preimplantation genetic testing: where we are today. *IJMS* 2020; **21**:4381.
- He Z, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachusetts, United States, 2015.
- Iwase A, Nakamura T, Osuka S, Takikawa S, Goto M, Kikkawa F. Anti-Müllerian hormone as a marker of ovarian reserve: What have we learned, and what should we know? *Reprod Med Biol* 2016; **15**:127–136.
- Kan-Tor Y, Ben-Meir A, Buxboim A. Can deep learning automatically predict fetal heart pregnancy with almost perfect accuracy? [Review of *Can deep learning automatically predict fetal heart pregnancy with almost perfect accuracy?*]. *Hum Reprod* 2020; **35**:1473.
- Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, Sigaras A, Lavery S, Cooper LAD, Hickman C et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ Digit Med* 2019; **2**:21.
- Kirillova A, Lysenkov S, Farmakovskaya M, Kiseleva Y, Martazanova B, Mishieva N, Abubakirov A, Sukhikh G. Should we transfer poor quality embryos? *Fertil Res Pract* 2020; **6**:2.
- Kragh MF, Karstoft H. Embryo selection with artificial intelligence: how to evaluate and compare methods? *J Assist Reprod Genet* 2021; **38**:1675–1689.
- Kragh MF, Rimestad J, Berntsen J, Karstoft H. Automatic grading of human blastocysts from time-lapse imaging. *Comput Biol Med* 2019; **115**:103494.
- Lassen JT, Kragh MF, Rimestad J, Johansen MN, Berntsen J. Development and validation of deep learning based embryo selection across multiple days of transfer. *arXiv preprint arXiv:2210.02120*, 2022.
- Lessey BA, Young SL. What exactly is endometrial receptivity? *Fertil Steril* 2019; **111**:611–617.
- Liao Q, Zhang Q, Feng X, Huang H, Xu H, Tian B, Liu J, Yu Q, Guo N, Liu Q et al. Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring. *Commun Biol* 2021; **4**:415.
- Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020; **2**:56–67.
- Martin JR, Mahutte NG, Arici A, Sakkas D. Impact of duration and dose of gonadotrophins on IVF outcomes. *Reprod Biomed Online* 2006; **13**:645–650.
- Meseguer M, Herrero J, Tejera A, Hilligsøe KM, Ramsing NB, Remohí J. The use of morphokinetics as a predictor of embryo implantation. *Hum Reprod* 2011; **26**:2658–2671.
- Mushtaq R, Pundir J, Achilli C, Najji O, Khalaf Y, El-Toukhy T. Effect of male body mass index on assisted reproduction treatment outcome: an updated systematic review and meta-analysis. *Reprod Biomed Online* 2018; **36**:459–471.
- Oron G, Son WY, Buckett W, Tulandi T, Holzer H. The association between embryo quality and perinatal outcome of singletons born after single embryo transfers: a pilot study. *Hum Reprod* 2014; **29**:1444–1451.
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al. PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*, Vol. **32**. Red Hook, NY, United States: Curran Associates, Inc., 2019, 8024–8035. <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachusetts, United States, 2015.
- Sanders KD, Silvestri G, Gordon T, Griffin DK. Analysis of IVF live birth outcomes with and without preimplantation genetic testing for aneuploidy (PGT-A): UK Human Fertilisation and Embryology Authority data collection 2016–2018. *J Assist Reprod Genet* 2021; **38**:3277–3285.
- Sawada Y, Sato T, Nagaya M, Saito C, Yoshihara H, Banno C, Matsumoto Y, Matsuda Y, Yoshikai K, Sawada T et al. Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth. *Reprod Biomed Online* 2021; **43**:843–852.
- Song H, Kim M, Park D, Shin Y, Lee J-G. Learning from noisy labels with deep neural networks: a survey. *IEEE transactions on neural networks and learning systems*, PP, 10.1109/TNNLS.2022.3152527. 2022. Advance online publication. <https://doi.org/10.1109/TNNLS.2022.3152527>.
- Stadtmauer L, Dittkoff EC, Session D, Kelly A. High dosages of gonadotropins are associated with poor pregnancy outcomes after in vitro fertilization-embryo transfer. *Fertil Steril* 1994; **61**:1058–1064.
- Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2015, 4489–4497.
- Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Hum Reprod* 2019; **34**:1011–1018.
- Tran D, Cooke S, Illingworth PJ, Gardner DK. Reply: Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer [Review of Reply: Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer]. *Hum Reprod* 2020a; **35**:483.
- Tran D, Cooke S, Illingworth PJ, Gardner DK. Reply: Can deep learning automatically predict fetal heart pregnancy with almost perfect accuracy? [Review of Reply: Can deep learning automatically predict fetal heart pregnancy with almost perfect accuracy?]. *Hum Reprod* 2020b; **35**:1474.
- Veiga E, Olmedo C, Sánchez L, Fernández M, Mauri A, Ferrer E, Ortiz N. Recalculating the staff required to run a modern assisted reproductive technology laboratory. *Hum Reprod* 2022; **37**:1774–1785.
- VerMilyea M, Hall JMM, Diakiw SM, Johnston A, Nguyen T, Perugini D, Miller A, Picou A, Murphy AP, Perugini M. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. *Hum Reprod* 2020; **35**:770–784.

- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J et al; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods* 2020;**17**:261–272.
- Wong CC, Loewke KE, Bossert NL, Behr B, De Jonge CJ, Baer TM, Reijo Pera RA. Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nat Biotechnol* 2010;**28**:1115–1121.
- Yang L, Peavey M, Kaskar K, Chappell N, Zhu L, Devlin D, Valdes C, Schutt A, Woodard T, Zarutskie P et al. Development of a dynamic machine learning algorithm to predict clinical pregnancy and live birth rate with embryo morphokinetics. *Field Staff Reports/UFSI* 2022;**3**:116–123.
- Yih MC, Spandorfer SD, Rosenwaks Z. Egg production predicts a doubling of in vitro fertilization pregnancy rates even within defined age and ovarian reserve categories. *Fertil Steril* 2005;**83**:24–29.
- Zabari N, Kan-Tor Y, Or Y, Shoham Z, Shofaro Y, Richter D, Har-Vardi I, Ben-Meir A, Srebnik N, Buxboim A. Delineating the heterogeneity of preimplantation development via unsupervised clustering of embryo candidates for transfer using automated, accurate and standardized morphokinetic annotation. *medRxiv preprint* 2022.03.29.22273137. <https://doi.org/10.1101/2022.03.29.22273137>.
- Zaninovic N, Rosenwaks Z. Artificial intelligence in human in vitro fertilization and embryology. *Fertil Steril* 2020;**114**:914–920.



QUALITY, INNOVATION, AND SERVICE—IT'S ALL AT THE CENTER OF EVERYTHING WE DO.

From developing assisted reproductive technologies that maximize performance, like the first ART media and cultures, to expertise that streamlines productivity, FUJIFILM Irvine Scientific brings together decades of industry expertise with a powerhouse of innovation, turning opportunities into realities. Together, we're working to support healthy futures—from retrieval to realization.

ALL IN FOR LIFE.
irvinesci.com/ALLIN

FUJIFILM
Value from Innovation

 **IrvineScientific**