scientific reports

Check for updates

OPEN

# Deep-learning model for embryo selection using time-lapse imaging of matched high-quality embryos

Lisa Boucret[1]✉, Floris Chabrun[2,4], Magalie Boguenet[1,4], Pascal Reynier[2,4], Pierre-Emmanuel Bouet[3] & Pascale May-Panloup[1,4]

**Time-lapse imaging and deep-learning algorithms are promising tools to assess the most viable embryos and improve embryo selection in IVF laboratories. Here, we developed and validated a deep learning model based on self-supervised contrastive learning. The model was developed with a new approach based on matched KID (Known Implantation Data) embryos derived from the same cohort of a stimulation cycle, both judged to be of good quality according to classical morphological criteria and morphokinetics, transferred fresh or frozen, but with a different implantation fate (clinical pregnancy vs. failure of implantation). We used self-supervised contrastive learning to train convolutional neural networks to ensure an unbiased and comprehensive learning of the morphokinetics features of the embryos, followed by a Siamese neural network fine-tuning and an XGBoost final prediction model to prevent overfitting. 1580 embryo videos of 460 patients were included between January 2020 and February 2023. With the knowledge of the implantation outcome of a previous transfer of an embryo derived from the same stimulation cycle, this model could predict the pregnancy outcome of the subsequent transfer with an AUC of 0.57. Without any knowledge of transfer history, the model achieved a satisfactory performance in predicting implantation (AUC = 0.64). This model could be considered as an adjunct tool for biologists to better select embryos and reduce the number of useless transfers per patient, when a cohort with several embryos classified as good quality by classical criteria is obtained.**

**Keywords** Embryo morphokinetics, Artificial intelligence, Time-lapse, Machine learning, Deep learning, Implantation

Since the first human baby resulting from in vitro fertilization (IVF) in 1978, more than 12 million children have been born after assisted reproduction technology (ART) treatment worldwide[1] and over 3.5 million cycles are performed every year. Despite this, IVF's success remains relatively low, leading to high financial and emotional costs. Embryo selection represents a critical step in the prognosis of the IVF treatment. Embryo morphology is traditionally used to assess embryo quality and to select embryos for transfer. However, under the classical approach, embryos are observed only at discrete observational time points, and the number of such usable conventional grading criteria is limited. Recently, the introduction of the time-lapse system (TLS) in IVF laboratories has enabled continuous monitoring of embryo development without disturbing culture conditions by removing embryos from the incubator. By combining a panel of morphokinetic parameters, several algorithms have been developed to facilitate the selection or the deselection of embryos for transfer[2–4]. Nevertheless, this semi-automatic approach is time-consuming and is limited by an inter- and intra-observer variability, due to the subjectivity of manual annotations. Additionally, a Cochrane review concluded that there is currently insufficient good-quality evidence of differences in live births to choose between TLS, with or without embryo selection software, and conventional incubation[5]. In this context, the increasing access to "big data" and artificial intelligence (AI) in reproduction medicine has generated new hope for optimization of the use of time-lapse for embryo selection. Deep learning algorithms, in particular convolutional neural networks (CNNs) have recently been used to directly predict embryo viability and quality[6]implantation[7–9] or live birth[10–13] by analyzing the raw time-lapse videos without the need for annotated parameters. This automated approach could be easily integrated into the workflow of a busy IVF laboratory.

[1]Reproductive Biology Laboratory, Angers University Hospital, Angers 49000, France. [2]Biochemistry and Molecular Biology Laboratory, University Hospital, Angers 49000, France. [3]Department of Gynecology and Obstetrics, Angers University Hospital, Angers 49000, France. [4]Mitolab, MitoVasc Institute, CNRS 6015, Inserm U1083, Angers University, Angers 49000, France. ✉email: liboucret@chu-angers.fr

In this study, we propose a deep-learning model built from a dataset obtained from matched KID (Known Implantation Data) embryos derived from the same cohort of a stimulation cycle, but with a different implantation fate. Only cycles with embryo freezing and several transfers with different outcomes (with and without implantation) are included, in order, on the one hand, to overcome individual patient characteristics, and on the other hand, to only include embryos judged to be of good quality by conventional morphological and kinetic criteria. We aimed to build a model that could provide additional information to those used in daily practice (conventional morphology and morphokinetics), independently of patient and cycle characteristics, and could potentially detect subtle differences between embryos from the same cycle that were judged to be similar according to these traditional methods. Therefore, the purpose of the current study is to evaluate the reliability of this deep-learning model, and to determine whether the AI model represents an added value in our daily practice of embryo selection.

## Materials and methods

### Study design and population

This research was a retrospective observational study carried out of IVF cycles peformed in the reproductive medicine center at Angers hospital (France) between January 2020 and February 2023. We included women from 18 to 43 years old (French standard age requirements for IVF) who obtained embryos cultured in an EmbryoScope+® time-lapse incubator and for whom the stimulation cycle resulted in several embryo transfers (fresh and/or frozen).

Namely clinical pregnancy (positive known implantation data, KIDp) or implantation failure (negative known implantation data, KIDn).

### Ovarian stimulation and laboratory procedures

Pituitary suppression was achieved using either an antagonist (Ganirelix; 0.25 mg daily), or a gonadotrophin-releasing hormone agonist (Triptorelin; 0.1 mg subcutaneously daily). Ovarian stimulation was performed using recombinant or urine-derived follicle stimulation hormone (FSH). Gonadotropin dose selection was based on patient characteristics, ovarian reserve biomarkers, and response to any previous ovarian stimulation cycle. Triggering criteria included a minimum of three follicles ≥ 17 mm and serum estradiol (E2) levels. Ultrasound-guided transvaginal oocyte retrieval was scheduled 36 h after ovulation triggering, which was performed by injection of 6500 IU choriogonadotropin alfa and/or 0.2 mg triptorelin.

Collected oocyte cumulus complexes were washed with G-MOPS™ PLUS (Vitrolife, Sweden), and placed for 1–3 h in FertiCult® IVF medium (FertiPro, Belgium) at 5% O2, 6% CO2, and 37 °C. Sperm preparation was performed using a standard gradient separation at 300 g for 20 min, followed by washing with FertiCult® IVF medium (FertiPro, Belgium) at 600 g for 10 min. Oocytes were inseminated by IVF or ICSI according to semen quality parameters and patient's history (failed or poor fertilization on a previous cycle). For conventional IVF, the cumulus-oocyte complexes were incubated with 100,000 motile spermatozoa in 1 mL of FertiCult® IVF medium and denuded the following day (19–20 h after insemination). For ICSI cycles, metaphase II (MII) oocytes were injected using a RI Integra 3™ Micromanipulator (Cooper Surgical Company, Trumbull, Conn., USA) after denudation with hyaluronidase (FertiPro, Belgium). Oocytes were then cultured in an EmbryoScope+® time-lapse incubator (Vitrolife, Sweden), in pre-equilibrated EmbryoSlides™ with a global culture medium (G-TL™, Vitrolife, Sweden) under a controlled atmosphere (5% O2, 6% CO2). Images were acquired automatically every 10 min in 11 focal planes with illumination from a single red LED (635 nm) until use. Embryo development was assessed with the EmbryoViewer software (Vitrolife, Sweden). Fertilization was checked ~ 19 h post insemination or injection. Abnormally fertilized oocytes (1 or 3 or more pronuclei) were excluded from further consideration. Number of cells, fragmentation level, symmetry among blastomeres and compaction degree were evaluated on days 2 and 3 of development, using the BLEFCO classification[14]. According to this classification, embryos that were ≥ 4.1.2. or 4.2.1. at day 2 and ≥ 8.1.2. or 8.2.1. at day 3 were deemed good grade, other types were deemed poor grade. Blastocysts were assessed according to the Gardner and Schoolcraft classification[15]. Based on these criteria, we defined good-quality blastocysts as follows: expansion grade ≥ 3, inner cell mass (ICM) grade ≥ B, and trophectoderm grade ≥ B on day 5. Any combination of ICM or trophectoderm quality grading of "C", or embryos with developmental stage graded as early blastocyst or below were classified as poor quality. Morphokinetic parameters were manually annotated according to published guidelines[16] and included: time to syngamy (tPNf), time to two (t2), three (t3), four (t4), five (t5), and eight (t8) cells, as well as time to blastocyst (tB). Embryos were then scored from 1 to 5 by the KIDScore D3 v1.2 algorithm[4] and from 1 to 9.9 by the KIDScore D5 v3.1 algorithm[17].

Embryos were selected for transfer or freezing according to the result of the KIDScore™ Day 3 or Day 5, and the conventional grading criteria. Other parameters such as multinucleation, direct or reverse cleavage, and blastocyst collapse were also monitored to further classify embryos. Embryos with abnormal (direct or reverse) cleavage were discarded. According to age, cycle number and quality of embryos, one or two embryos were transferred under transabdominal ultrasound guidance at cleaved or blastocyst stage. Vitrification was performed at cleaved or blastocyst stage using closed CBS High Security Vitrification (HSV) straws (Cryo Bio System, France) in combination with ethylene glycol, DMSO and sucrose as the cryoprotectants (Vit Kit-Freeze and Vit Kit-Thaw, Irvine Scientific, USA), as described previously[18].

Luteal phase support of fresh embryo transfers was provided with oral dydrogesterone (30 mg/day) and personalized in case of abnormal endometrial receptivity analysis or previous implantation failure. In case of frozen embryo transfers, endometrial preparation was conducted via artificial cycles with oral or transdermal estrogen, and the addition of intravaginal progesterone when endometrial thickness was ≥ 8mm[19]. Biological pregnancy was confirmed 14 days after oocyte retrieval with a serum β-HCG level above 100 IU/L. Clinical pregnancy was defined as the presence of at least one fetal heartbeat on ultrasound 5 weeks after embryo transfer.

Double embryo transfers (DET) that resulted in a single gestational sac were excluded. During the inclusion period, there was no significant difference in pregnancy outcomes between fresh and frozen embryo transfers. We matched embryos obtained from a same cycle in pairs according to their clinical outcome.

## Data collection and image preprocessing

Raw videos were exported using the EmbryoViewer software (Vitrolife, Sweden) and then processed in Python. Data preprocessing was applied to convert the videos into usable images, given that raw videos could not be exploited directly. First, the initial resolution ($400 \times 400$) was too large to be directly processed as sequences of images with the available computational resources. Thus, we cropped all images by restricting them to the view around the embryo. We also discarded all frames of poor quality, containing artefactual visual defects, such as bubbles or low luminosity and hindering the visibility of the embryo, and all frames without embryo (due to an asynchronous time of transfer and freezing for the same embryo cohort).

To carry out this step, we trained a YOLO v6 object detection deep learning algorithm to detect and locate the embryos in raw images[20]. We first curated a random subset of 2000 images of embryos at various stages. Using color binarization and thresholding, we semi-automatically located the embryos on these images, before manually reviewing these annotations. Finally, after image augmentation (random rotation, flipping, cropping, affine and elastic transformations, noise, blur), YOLO was trained to detect and locate embryos. After training, YOLO was used to preprocess all videos into embryo images: images in which an embryo was detected and located were kept and cropped/padded with black to $362 \times 362$, centered on the embryo. We manually reviewed all images to ensure the quality of our dataset. Images in which an embryo could be visually identified but was undetected by YOLO due to visual artefacts were discarded.

## Pre-training: extracting embryo morphology features

We used self-supervised contrastive learning, namely the SimCLR architecture, to pre-train a deep learning model (encoder) to learn to map (encode) embryo images into feature vectors summarizing static morphological features[21]. This step was performed on videos of embryos that fulfilled the inclusion criteria summarized in Table 1. In contrast to the original SimCLR architecture, we used a more conservative image augmentation pipeline, limited to random rotation and flipping followed by centered cropping, minor contrast, and saturation jittering. However, instead of using two augmentations of the same original image, each positive pair was composed of two different images of the same embryo taken 10 min to $<1$ h apart. This was performed to ensure that the model could learn to map both the stage of the embryo but also its intrinsic morphological characteristics into the final feature vector. Two architectures were trained and compared, namely VGG16 [22] and ResNet18.

During this step, the encoders were trained on all images of embryos selected for transfer or freezing, regardless of the transfer outcome. To validate the model during and after pre-training, we used a publicly available external dataset[23,24]. The dataset comprised 302,134 frames of videos accompanied by the annotations of 16 cellular events, from the extrusion of the second polar body to the hatched blastocyst stage. Annotation of these timings was performed and manually checked by experienced embryologists using the definition of key events proposed by Ciray et al.[16]: tPB2 (extrusion of the second polar body), tPNa (pronuclei appearance), tPNf (pronuclei fading), t2, t3, t4, t5, t6, t7, t8, t9+, tM (end of compaction), tSB (start of blastulation), tB (full blastocyst), tEB (expanded blastocyst), and finally tHB (hatched blastocyst). Since the number of images per event in the dataset was imbalanced, we randomly sampled 99 images per stage of development to ensure a balanced external validation dataset, thus including a total of 1,584 images.

After each epoch and during the final validation step, we used the deep learning pre-trained model to encode the images of embryos into feature vectors, and then trained an XGBoost model to predict the stage of the embryo based on these feature vectors. We used mean one-versus-one Area Under the Receiver Operating Characteristic Curve (ROC-AUC) and F1-score after 5-fold cross-validation as the metric to test how much information from the embryos images were kept and encoded by the pre-trained model into the feature vector.

## Fine-tuning: extracting morphokinetic features

After pre-training, we used supervised one-shot learning (Siamese network) to fine-tune the encoders to predict the pregnancy outcome of the embryo transfer, namely clinical pregnancy (positive known implantation data, KIDp) or implantation failure (negative known implantation data, KIDn). For fine-tuning, the training dataset was limited to cycles with at least three transfers, including at least one positive and one negative. This was done to ensure that, in the same cohort, each embryo could be matched both to an embryo with a similar outcome

| Inclusion criteria | Exclusion criteria |
|---|---|
| Age 18–43<br>Cycles involving multiple embryo transfers<br>Known implantation data<br>Matched embryos from the same cycle with similar (KIDn/KIDn or KIDp/KIDp) or different outcome (KIDn/KIDp)<br>High grade embryos<br>Day 2 : ≥ 4.1.2. or 4.2.1. (Blefco classification<br>Day 3 : ≥ 8.1.2. or 8.2.1. (Blefco classification)<br>Day 5 : ≥ 3BB (Gardner classification) | Objection to processing of personal data<br>DET with a single gestational sac<br>Artefactual image defects |

**Table 1**. Inclusion and exclusion criteria of the study. *KIDp* positive known implantation data, *KIDn* negative known implantation data, *DET* double embryo transfer.

and to an embryo with a different one, in order to prevent overfitting. Random stratified partitioning per patient was used to separate videos of embryos into a training set (80%) and a validation set (20%). No embryo could be found both in the training and in the validation set. All videos unsuitable for fine-tuning (i.e. less than three usable embryos or lack of a positive or a negative outcome) were assigned to the validation set.

Each training step was performed as follows (Fig. 1): (1) for a same cycle, two embryo videos were randomly selected, and a sequence of N random images of each of these videos was randomly selected, ensuring that each pair of images was from the same cycle and at the same time point; (2) each image was encoded using the previously pre-trained model into a "static" feature vector; (3) each of the two sequences was input to a long short-term memory (LSTM) model, which output one "dynamic" feature vector for each embryo of the same patient, summarizing their morphokinetic features; (4) the morphokinetic feature vectors were compared and the model was penalized into predicting similar morphokinetic feature vectors(low Euclidean distance) for

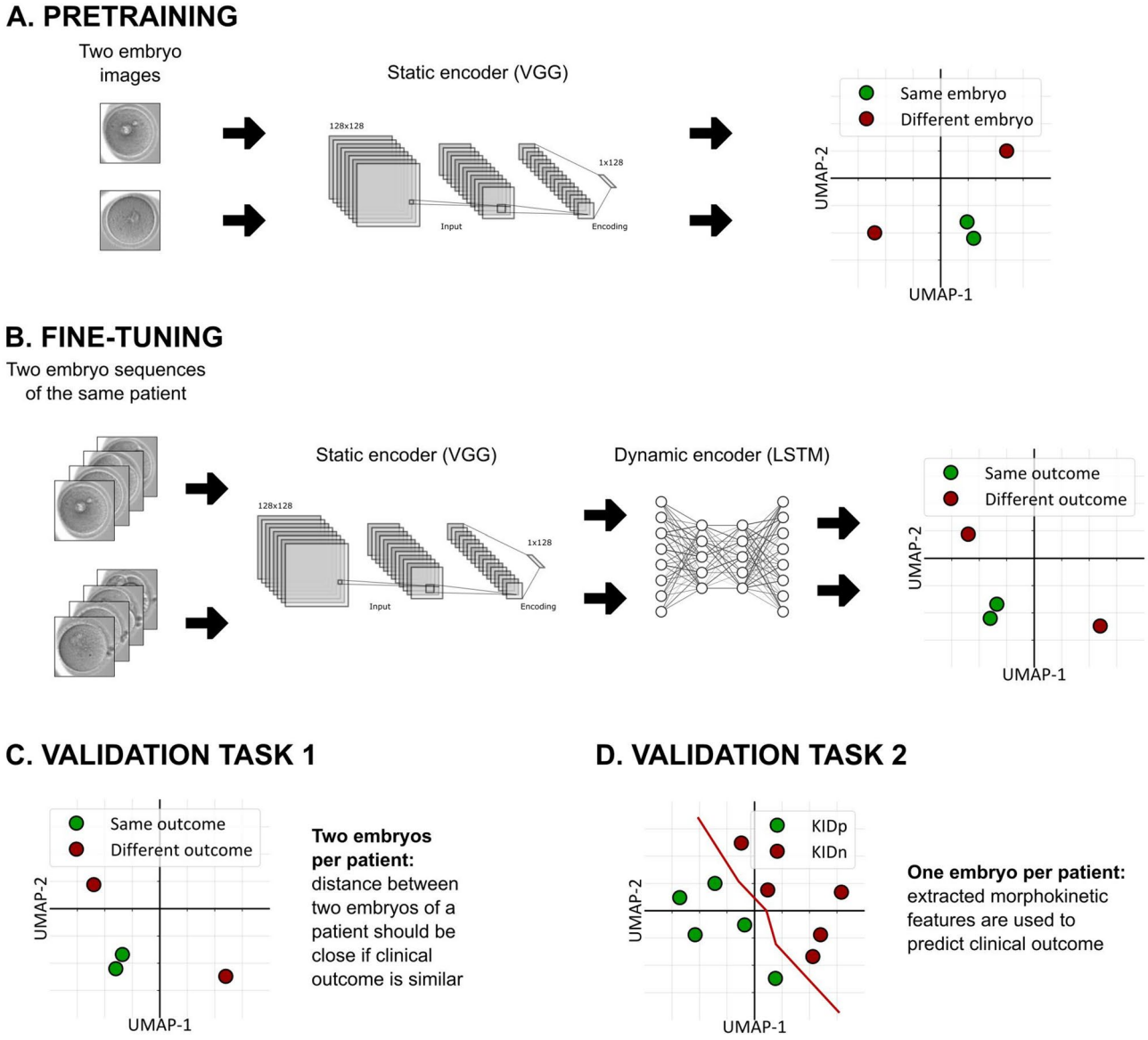


**Fig. 1**. Training and validation steps of the deep learning models used for implantation prediction. (**A**) Self-supervised contrastive learning is used to pre-train the static encoder to extract morphological features from embryo images. (**B**) One-shot learning is used to train the dynamic encoder to extract morphokinetic features correlated to the clinical pregnancy outcome. The model is trained on paired embryo sequences, and the Euclidean distance between the paired output morphokinetic feature vectors is compared to the clinical pregnancy outcome pair. (**C**) Morphokinetic features from two embryos of the same stimulation cycle are compared to predict whether that of the clinical pregnancy outcome is similar (KIDn/KIDn or KIDp/KIDp) or different (KIDn/KIDp). (**D**) Features from one embryo per patient are used to predict the clinical outcome of the attempt. *VGG* visual geometry group, *LSTM* long short-term memory, *KIDp* positive known implantation data, *KIDn* negative known implantation data.

videos of embryos with the same clinical outcome (KIDp/KIDp or KIDn/KIDn), and dissimilar (high Euclidean distance) for embryos with a different clinical outcome (KIDn/KIDp or KIDp/KIDn).

Several hyperparameters were compared during fine-tuning, namely: input size of images (64×64 or 128×128), training or freezing the weights of the pre-trained static encoder model, number of frames per 48 h (6, 12, 24, or 48), learning rate, batch size and number of epochs.

After fine-tuning, the models' predictions were tested using two validation tasks. The first validation task consisted of predicting the clinical pregnancy outcome by knowing the outcome of a previous transfer performed with an embryo from the same cohort. For this task, we included cycles with at least two paired embryos (with identical or different outcomes). After encoding both paired embryo videos by the fine-tuned model, the Euclidean distance between the two encodings was used as a metric to determine whether the two embryos of the pair should be associated with the same outcome. We performed a 1000-iteration bootstrap analysis; for each bootstrap, we used the samples resampled with replacement to determine the ROC-AUC and optimal Euclidean distance threshold using the Youden index. Then, we used the samples left out of the bootstrap to determine F1-score, sensitivity and specificity at the determined threshold. Finally, results from all iterations were aggregated to compute 95% confidence intervals.

The second validation task was to predict the clinical pregnancy outcome of a transfer without knowing the outcome of a previous transfer. For this task, we included one embryo for each cycle. Based on the encodings output by the fine-tuned model, we trained a simpler machine learning model, namely XGBoost with 5-fold cross-validation repeated over 200 iterations, to predict the clinical outcome based on these morphokinetic feature vectors. For each model, we computed F1-score, ROC-AUC, sensitivity and specificity, and merged the means of all 200 iterations to compute 95% confidence intervals. In parallel, we also ran a permutation test with 200 random permutations for each 5-fold cross-validated model to assess the significance of the model's performance compared to random.

### Statistical analyses and model performance

The performance of the deep learning model was assessed using AUC of ROC curve generated by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across all possible thresholding values. AUC score ranges from 0 to 1, where 0.5 indicates random classifier, and 1 indicates perfect predictive performance. AUC values were interpreted according to the classification as described elsewhere[25]and 95% confidence intervals were compared to an AUC of 0.5 to assess performance against a random model. We also used the weighted F1-score, sensitivity and specificity, using a threshold of 0.5 (50%). 95% confidence intervals were computed over 1000-iteration bootstrapping, or using 5-fold cross-validation repeated over 200 iterations, by selecting the 2.5th and 97.5th percentiles over repeats.

### Computational tools

We used Python 3.11.5, Pillow 9.3.0, scikit-learn 1.3.2, scikit-image 0.24.0, pytorch 2.1.1, imgaug 0.4.0 to perform analyses.

## Results

A total of 3419 videos of 504 patients were analyzed between January 2020 and February 2023 (Fig. 2). After discarding embryos due to poor morphological quality, 1580 embryos from 460 patients were available and used for pre-training. Demographic characteristics and IVF parameters of cycles are described in Table 2. After excluding embryos with unusable data (i.e. embryos frozen but never transferred, DET with only one single gestational sac, image artefacts preventing image sequence generation), 829 embryos from 374 patients were still available. Among these, 209 embryos from 62 patients were used to fine-tune the model, and 620 from 312 patients were used to define the validation cohort. The first validation task (predicting the clinical pregnancy outcome of an embryo with the knowledge of a matched embryo from the same cohort) was applicable to all cycles with at least two embryos with identical or different outcomes. One hundred seventy-four patients from the validation set matched these criteria and were selected for this task. To ensure unbiased results, only one pair of embryos was selected for each patient, prioritizing KIDp/KIDn and KIDp/KIDp pairs when available. The second validation task (predicting the pregnancy outcome without knowing the previous transfer outcome) could be applied to all 312 patients included in the validation set. To prevent bias, we only included one embryo per patient, resulting in 312 embryos for 312 patients.

### Statistical analyses and model performance

After pre-training, the encoder was validated on an external dataset of 1,296 embryo images evenly divided into 16 classes, based on annotations checked manually[23]. We trained XGBoost models to determine the discriminant power of these encodings to discriminate embryos based on their stage in a 1-*versus*-1 stage setting, by computing the mean weighted F1-score and ROC-AUC over a five-fold cross-validation supervised training. Overall, XGBoost models were able to discriminate between two stages with a median F1-score of 0.92 throughout all 1-*versus*-1 comparisons, with interquartile range 0.85–0.96 (Table 3). Median ROC-AUC for all 1-*versus*-1 comparisons was 0.97 (IQR: 0.93–0.99) (see Supplementary Table S1).

Interestingly, using a ResNet18 architecture instead of the VGG architecture slightly improved the median F1-scores and ROC-AUC for the pre-training validation task (F1-score: median of 0.94, IQR: 0.88–0.98; ROC-AUC: median of 0.98, IQR 0.95-1.00) but did not improve the pregnancy outcome prediction after fine-tuning.

### Dynamic encoding and clinical pregnancy prediction

After fine-tuning, the encoder which performed the best was VGG16, after training for 11 epochs at a rate of 1 frame per 4 h. For the first validation task (174 patients for whom videos of KID embryos pairs were available),

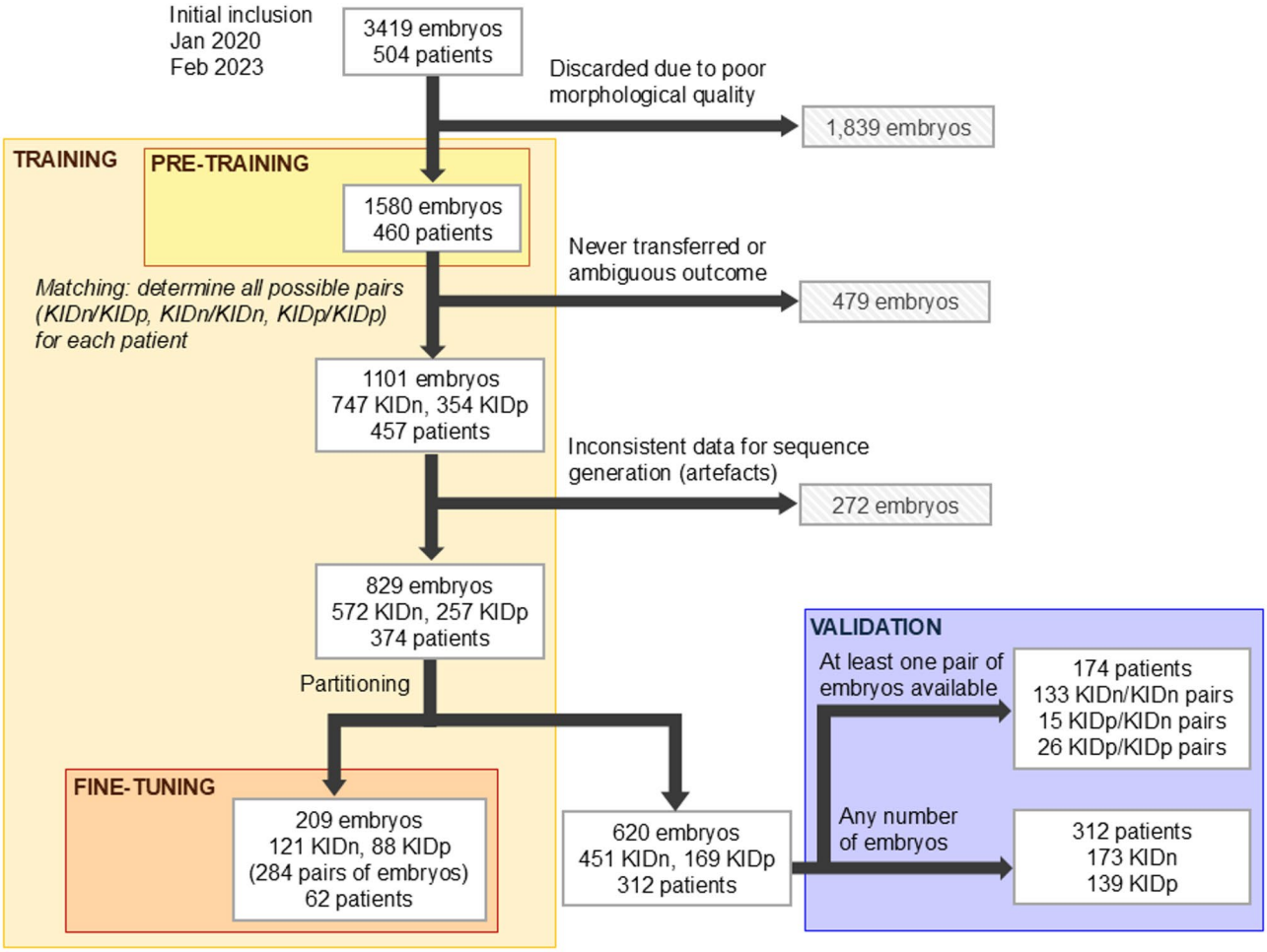


**Fig. 2**. Description of the cohort and samples used for pre-training, fine-tuning, and model validation.

we used 1000-iteration bootstrapping to determine the discriminant power of the Euclidean distance between the morphokinetic encodings of the two embryo videos, to determine the outcome of a transfer knowing the outcome of a previous transfer. The distance between two embryos yielded a mean F1-score of 0.14 (95% CI: 0-0.28), with a mean ROC-AUC of 0.57 (95% CI: 0.41–0.74). Mean sensitivity and specificity were 48.3% (95% CI: 0-100%) and 54.8% (95% CI: 8.8–89.8%), respectively.

The second validation task obtained a mean F1-score of 0.55 (95% CI: 0.51–0.60) with XGBoost models trained using 5-fold cross-validation to discriminate the clinical pregnancy outcome of 312 patients (173 with a negative outcome, 139 with a positive outcome). A permutation test, repeated over 200 iterations, showed that the F1-score was significantly different from random ($p = 0.02$). Mean ROC-AUC was 0.64 (95% CI: 0.60–0.68). Mean sensitivity and specificity were 53.9% (95% CI: 48.7–59.8%) and 68.1% (95% CI: 63.1–73.3%), respectively. Calibration analysis showed a Brier score of 0.30 (95% CI: 0.25–0.36) (Fig. 3).

## Feature exploration

The unsupervised analysis (Fig. 4), including 312 embryos from the second validation dataset, showed that there was no clear clustering of KIDp or KIDn embryos. When selecting a high number of clusters ($n = 32$), we obtained 7 clusters (22%) with 3 or more embryos and more than two-thirds with a positive outcome; and 13 clusters (41%) with 3 or more embryos and more than two-thirds with a negative outcome.

## Discussion

In this study, we developed and validated a deep learning model based on self-supervised contrastive learning. This model was built from a dataset obtained from matched KID embryos derived from the same cohort of a stimulation cycle but with a different implantation fate. The model achieved to predict implantation of morphologically good-quality embryos with an AUC of 0.64 (95% CI: 0.60–0.68). In this way, this model could be used in clinical routines to assist embryologists in embryo assessment and represents a promising tool for predicting the success of an embryo transfer. When adding the knowledge of the pregnancy outcome of a previous transfer performed with an embryo derived from the same stimulation cycle, the model could predict the pregnancy outcome of a subsequent transfer with an AUC of 0.57 only. We hypothesize that there is no single morphokinetic pattern indicating the quality of an embryo. This assumption is supported by the results of the

| | Pre-training database | Fine-tuning database | Validation database | |
|---|---|---|---|---|
| | | | Task 1 | Task 2 |
| | 1580 embryos | 209 embryos | 348 embryos | 312 embryos |
| **Cycle characteristics** | | | | |
| Oocyte age | 32.7 ± 4.6 | 32.2 ± 4.5 | 33.7 ± 4.9 | 33.1 ± 4.8 |
| Tobacco use | | | | |
| - Yes | 11.7% | 12.9% | 12.1% | 10.9% |
| - No | 85% | 82.3% | 86.8% | 86.2% |
| - Unknown | 3.3% | 4.8% | 1.1% | 2.9% |
| Woman BMI (kg/m$^2$) | 24.2 ± 4.9 | 23.9 ± 4.6 | 24.2 ± 5.0 | 24.3 ± 4.8 |
| Paternal age | 34.9 ± 5.4 | 34.6 ± 5.3 | 35.9 ± 5.3 | 35.2 ± 5.5 |
| Insemination method | | | | |
| - Conventional IVF | 53.7% | 58.1% | 50.0% | 50.3% |
| - ICSI | 46.3% | 41.9% | 50.0% | 49.7% |
| **Embryo parameters** | | | | |
| Stage | | | | |
| - Cleavage stage | 74.9% | 70.3% | 69.8% | 74.4% |
| - Blastocyst | 25.1% | 29.7% | 30.2% | 25.6% |
| Cryopreservation status | | | | |
| - Fresh | 33.0% | 26.3% | 40.8% | 43.3% |
| - Frozen | 67.0% | 73.7% | 59.2% | 56.7% |
| Positive outcome | 32.2% | 42.1% | 19.3% | 44.6% |

**Table 2**. Description of cycle characteristics and embryo parameters in the pre-training, fine-tuning and validation (Task 1 and 2) phases. Values are shown as mean ± SD or %.

| | tPNa | tPNf | t2 | t3 | t4 | t5 | t6 | t7 | t8 | t9p | tM | tSB | tB | tEB | tHB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| tPB2 | 0.98 | 0.98 | 0.95 | 0.94 | 0.98 | 0.82 | 0.92 | 0.98 | 0.91 | 0.87 | 0.96 | 0.97 | 0.93 | 0.99 | 0.84 |
| tPNa | | 0.89 | 0.77 | 0.92 | 0.92 | 0.98 | 0.96 | 0.83 | 0.96 | 0.94 | 0.9 | 0.77 | 0.93 | 0.97 | 0.98 |
| tPNf | | | 0.85 | 0.93 | 0.93 | 0.96 | 0.94 | 0.89 | 0.94 | 0.93 | 0.94 | 0.8 | 0.91 | 0.96 | 0.97 |
| t2 | | | | 0.89 | 0.78 | 0.98 | 0.91 | 0.75 | 0.88 | 0.93 | 0.81 | 0.8 | 0.84 | 0.96 | 0.99 |
| t3 | | | | | 0.89 | 0.94 | 0.77 | 0.89 | 0.87 | 0.84 | 0.81 | 0.92 | 0.73 | 0.95 | 0.94 |
| t4 | | | | | | 0.98 | 0.88 | 0.86 | 0.86 | 0.95 | 0.81 | 0.87 | 0.83 | 0.97 | 0.98 |
| t5 | | | | | | | 0.95 | 0.96 | 0.91 | 0.82 | 0.94 | 0.96 | 0.92 | 0.97 | 0.78 |
| t6 | | | | | | | | 0.88 | 0.79 | 0.68 | 0.78 | 0.89 | 0.71 | 0.96 | 0.92 |
| t7 | | | | | | | | | 0.94 | 0.96 | 0.88 | 0.81 | 0.87 | 0.96 | 0.99 |
| t8 | | | | | | | | | | 0.73 | 0.83 | 0.91 | 0.72 | 0.96 | 0.94 |
| t9p | | | | | | | | | | | 0.87 | 0.93 | 0.8 | 0.97 | 0.9 |
| tM | | | | | | | | | | | | 0.85 | 0.72 | 0.98 | 0.95 |
| tSB | | | | | | | | | | | | | 0.83 | 0.96 | 0.97 |
| tB | | | | | | | | | | | | | | 0.99 | 0.95 |
| tEB | | | | | | | | | | | | | | | 0.98 |

**Table 3**. Stage classification performance of the encoder model after pre-training, represented by weighted F1-score of 1-*versus*-1 discrimination. Timings of expected events: tPB2, second polar body extrusion; tPNa, pronuclei appearance; tPNf, pronuclei fading; t2-8, 2–8 cells; t9p, 9 cells or more; tM, end of compaction; tSB, start of blastulation; tB, full blastocyst; tEB, expanded blastocyst; tHB, hatched blastocyst.

unsupervised analysis, which shows no clear clustering between KIDp and KIDn embryos unless the number of clusters is increased.

Prioritizing embryos for transfer is a long-standing challenge in the field of IVF. Since the introduction of time-lapse microscopy in IVF laboratories, several algorithms have been developed which take into account the time at which a number of key morphological events occur[26]. Unfortunately, these classical models overlook numerous features whose relevance has not been previously investigated (like colour and granularity of cytoplasm, cell shape, and movement patterns…), and suffer from the subjectivity of manual annotations. Deep learning algorithms offer new perspectives since they can directly analyse the raw time-lapse sequence without assumptions about the significance of the information that embryologists focus on. One of the main concerns, however, is the lack of transparency and explainability, leaving embryologists to rely on a black-box model to

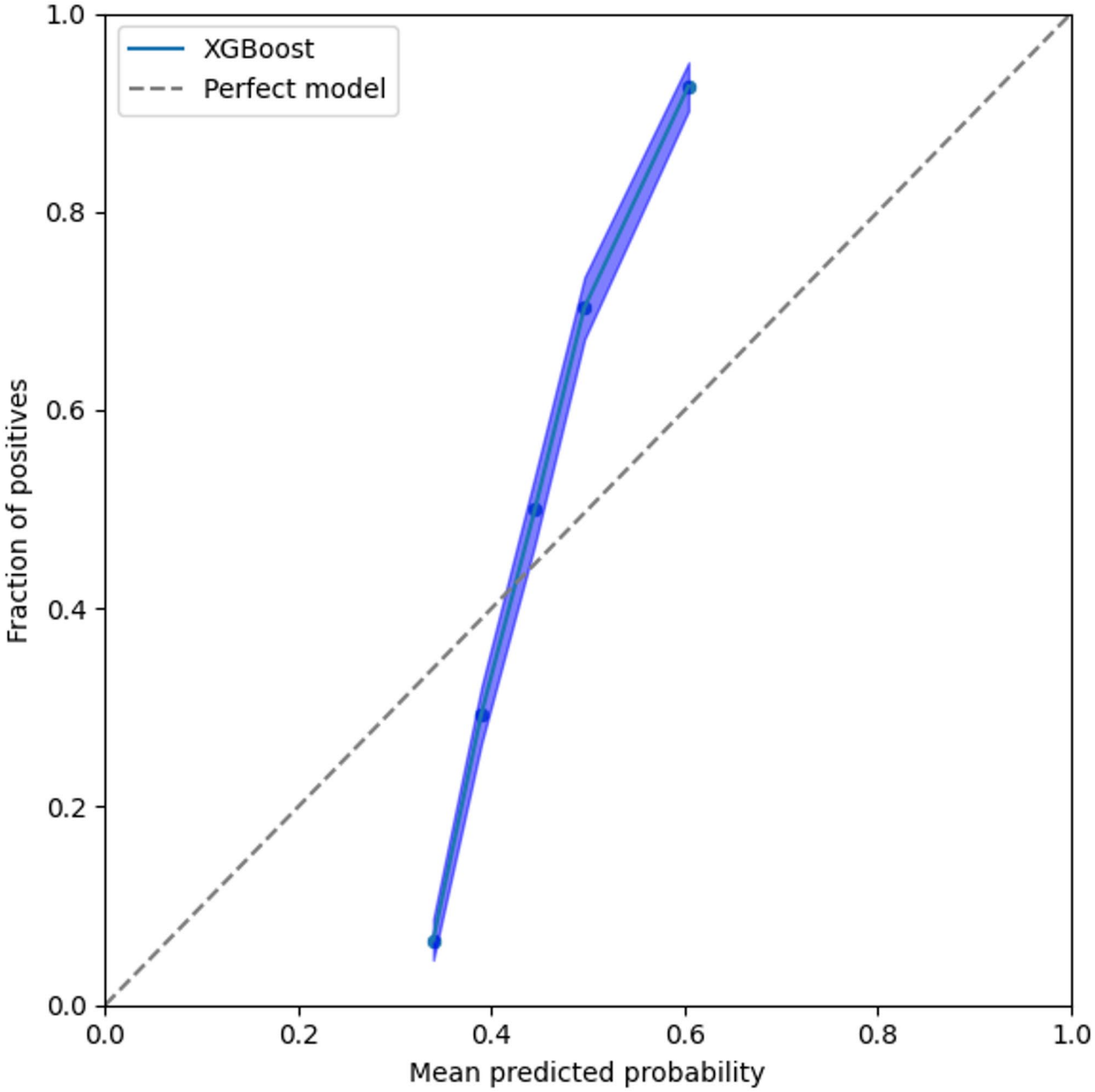


**Fig. 3**. Calibration plot of validation task 2: mean and 95% confidence interval of the actual fraction of positives according to the probability predicted by an XGBoost model, computed using 1000-iteration bootstrapping.

make clinical decisions. Despite this, this machine learning (ML) algorithm offers advantages over conventional methods for embryo selection. It offers new opportunities to better prioritize embryos for transfer in order to shorten the time to pregnancy[27] when a cohort with several embryos classified as good quality by classical morphokinetics is obtained. Furthermore, it saves valuable time in the embryo evaluation process. For example, a recent randomized controlled trial comparing deep learning algorithm (iDAScore) with standard morphology assessment found a 10-fold reduction in the time required for embryo evaluation[28].

In recent years, several ML algorithms have been developed in the field of IVF, with various names (STORK[29] IVY[9] FiTTE[30] iDAScore[31] ERICA[32] BlastAssist[33]…) and various outcomes such as embryo quality[27,29,34] blastocyst formation[12,29,35–37] ploidy status[27,38–40] implantation[7,9,40,41] clinical pregnancy[12] miscarriage[12] and live birth[11,12,42,43]. The algorithm described in this paper differs from previously published approaches in terms of the experimental design since it includes only cycles with good-quality embryos, and for which several consecutive transfers were performed. Although performances of different models trained with distinct datasets cannot be directly compared with each other, the AUC value of our model could be considered satisfactory compared with other models, especially because, as mentioned, this model enabled us to discriminate between embryos

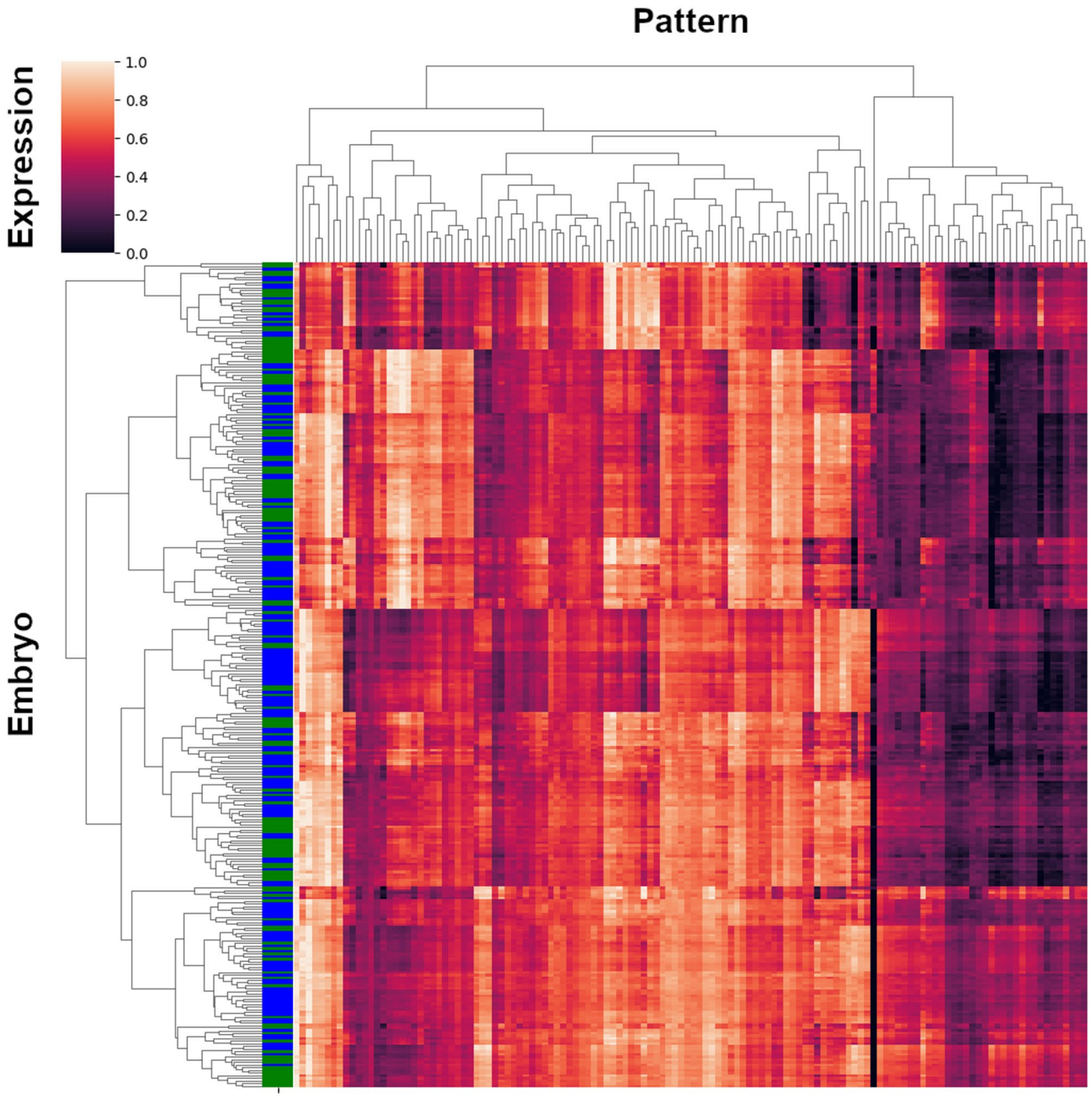


**Fig. 4**. Heatmap of the hierarchical clustering of morphokinetic features of embryos (columns) and clinical pregnancy outcome (rows). Embryos are depicted in green or blue according to their clinical outcome, respectively positive or negative.

of good quality (suitable for transfer or freezing according to classical morphokinetic criteria) and not in an exhaustive cohort of embryos of inconsistent quality. For example, a pioneer study obtained a surprisingly high AUC (0.93), but their dataset included 66% of discarded embryos[9]. Similarly, the first deep learning model developed by Berntsen et al.. (iDAScore v1.0) obtained an AUC of 0.95 for sorting the whole cohort of available embryos in relation to fetal heartbeat, but an AUC of 0.67 if only KID embryos were considered[7]. Following studies published AUCs ranging from 0.62 to 0.77 [41,44–47], moderately higher than those obtained with conventional embryo selection algorithms[26]. Indeed, some studies aimed to compare ML models with current embryo selection schemes based on morphology and/or morphokinetics. For example, Ueno et al.[31] compared the pregnancy prediction performance after single vitrified-warmed blastocyst transfer among 3 assessment methods: iDAScore v1.0 (automated embryo scoring system), KIDScore D5 v3 (annotation-dependent morphokinetic embryo scoring model) and Gardner criteria (traditional morphological grading model). The AUCs for the iDAScore, KIDScore and Gardner criteria were 0.70, 0.69 and 0.67 respectively and iDAScore AUC was significantly greater compared to other methods in the < 35 years age subgroup only.

Ehrlich et al.[48] developed a pseudo-label guided contrastive learning model in order to predict the KID status and compared it with KIDScore D3 system and a group of eight senior professionals. They found that their final ML model (AUC = 0.681) outperformed KIDScore D3 model (AUC = 0.582) and the group of human embryologists (mean AUC = 0.639) in general population but also across different age groups. Illingworth et al.[28] conducted the first randomized, double-blind noninferiority trial comparing deep learning-based embryo selection (iDAScore) with manual morphology-based assessment for single blastocyst transfer. The principal finding was that this study was not able to demonstrate noninferiority of deep learning for clinical pregnancy rate when compared to standard morphology used by embryologists. An overview of embryo selection through artificial intelligence compared to current ranking schemes based on morphology and/or morphokinetics is detailed in the literature[49–54] with description of advantages (data integration capabilities, automation, efficiency, potential for improved prediction, objectivity and reproducibility) and disadvantages (lack of transparency and interpretability, data quality and bias concerns, limited clinical validation, data security, ethical and regulatory concerns, cost and technical requirements), highlighting that further research is needed to improve embryo selection methods. While the model evaluated in this study achieved an AUC of 0.64, discriminative performance of ML algorithms is of key importance regarding their clinical application, particularly if applied for embryo deselection. A model with moderate predictive accuracy might inadvertently discard viable embryos, thereby reducing the overall likelihood of achieving a pregnancy. This is especially critical when such models are used not to rank embryos for transfer, but to exclude those deemed unlikely to result in pregnancy. Indeed, a low sensitivity model might fail to identify embryos with implantation potential, while a low specificity model might inappropriately prioritize embryos with poor prognosis. In both cases, the clinical outcome (cumulative pregnancy rate or time to pregnancy) could be negatively impacted.

An important consideration in the development of ML models applied to embryo selection is to know whether these models should integrate external factors as input data. This issue must be addressed when considering the generalizability of ML models trained on videos only, to heterogeneous patient populations from different clinics. Implantation is, in fact, multifactorial and depends not only on embryo quality but also on several clinical factors, including age, BMI, uterine receptivity, sperm quality, and stimulation procedures. It is plausible to imagine that including some relevant clinical parameters could improve the performance of this model. For example, Duval et al. found that training a hybrid model consisting of TLS videos and 31 clinical variables describing the patients and their IVF treatment significantly increased the AUC compared to algorithms that only analyzed videos (AUC = 0.73 *versus* 0.68 respectively)[44]. This is consistent with the results of Zou et al., who improved the predictive power of their models after adding clinical features to TLS parameters as input data[40]. However, published data are heterogeneous on this subject. Enatsu et al. found that the difference between the AUC of image-only and ensemble models was not statistically significant[30]. To further investigate the potential of ML models in a clinical setting, the choice of clinical variables to be included would be of great importance. Liu et al. ranked 103 patient couples' clinical features according to their ability to predict a live birth outcome and identified that 16 improved live birth prediction[11]. However, the most significant explanatory variables seem to differ substantially from one model to another. The so-called FiTTE AI system developed by Enatsu et al. revealed that after the blastocyst images, the best predictors of clinical pregnancy were age, pregnancy history, serum AMH, serum oestradiol, and progesterone at the time of embryo transfer[30]. The SHapley Additive exPlanations (SHAP) analysis of the hybrid model of Duval et al. showed that the most important features to predict pregnancy were video score, oocyte age, total gonadotrophin dose intake, number of oocytes, and embryos obtained and endometrium thickness[44]. Another study published by Blank et al.[41] found that gravidity and parity, age, and AMH levels were the most important predictive variables in the first two nodes of their random forest model (RFM). From the perspective of future studies, the model might also be improved by including other data sources like ploidy status, metabolic profiling and mitochondrial content. For example, the artificial neural network (ANN) models described by Bori et al. using blastocyst image analysis and proteomic profile of spent culture media such as concentrations of interleukin-6 and metalloproteinase-1 were able to predict live birth with excellent AUCs[55]. Finally, it should be mentioned that including some clinical factors does not a priori affect the ranking process itself, since these factors are constant for all the embryos derived from a same stimulation cycle, so this approach does not seem relevant in the context of this study.

This study included a large panel of patients, regardless of age, insemination method (IVF and ICSI), transfer protocol (fresh and cryopreserved), and transfer stage (day 2,3 or blastocyst). To investigate the generalization performance of the model across different patient demographics and clinical practices, it could be interesting to apply it in different subgroups, especially in different age groups. Different studies reported a higher AUC for older patients[43,44,46,48]. The most plausible explanation lies in a wider distribution of embryo quality for older women[46]. But as hypothesized by Ehrlich et al.[48]it cannot be ruled out either that infertility in younger patients is often due to non-embryonic causes, such as endometrium receptivity, resulting in noisy labels that can adversely affect the performance of the predictive model. Subgroup analyses of various studies[7,44,46] reported better predictive performance with fresh embryos than with cryopreserved embryos, probably because other factors (cryopreservation technique, endometrial preparation protocol) can have an impact on the likelihood of pregnancy. Concerning the stage of transfer, other studies validated AI models separately on cleavage stage embryo transfers[10,56] or blastocysts transfers[12]. Duval et al. found no statistical difference in AUC according to the day of transfer, suggesting that their AI model could detect key early phenomena[44]while Theilgaard Lassen et al. observed better performance for blastocyst-stage transfers compared with cleavage-stage ones, emphasizing that maximizing the amount of information available to AI algorithms can improve their predictive power[46].

To overcome the lack of transparency of AI systems, approaches using Class Activation Mapping (CAM) methodology could offer embryologists new insights into identifying embryo areas on which ML models focus to predict their outcomes. For example, Sawada et al. developed an attention map to visualize embryo features in focused regions associated with a live birth[42]. Unfortunately, no standard features were identified in the embryos

that could predict a live birth, even though there were many images in which high-focused areas existed around the zona pellucida. More specifically, the authors found no significant difference in the thickness of the zona pellucida between embryos that led to live birth and those that failed to implant and suggested that the neural network may have focused on the shape and density of the zona pellucida rather than on its thickness. More recently, heatmaps generated from the CNN model developed by Liu et al.[11] showed that trophectoderm-related features contributed more to live birth prediction when training included both blastocyst images and patient couple's clinical features, compared with training including blastocyst images only. These preliminary studies suggest that ML models likely detect a combination of known and previously unknown morphological features related to embryo quality[27] and highlight the need for further investigation to better characterize the relevant patterns and their biological significance.

Compared to other machine learning developed in the field, several strengths of this research can be highlighted. The algorithm is fully automated; its implementation in routine use does not need any additional manual intervention. A robust testing and validation process was employed to ensure the safety of the model. As deep neural networks are typically overparameterized and easily overfit to the training data, we applied a methodology (self-supervised learning and one-shot learning) to control the risk of overfitting and limit the bias in comparison to supervised learning methods only. This is demonstrated by the ability of our static encoder to accurately determine the stage of embryos without being explicitly trained for this task. In addition, this approach enables, during the pre-training step, the use of data of embryos for which the outcome is undetermined or ambiguous. Indeed, only the last simpler model, namely XGBoost, which is known for its reduced risk of overfitting compared to neural networks[57,58]is trained with supervised learning. Additionally, the training and the validation sets included embryos transferred at different times, reflecting the diversity of laboratory practices. Lastly, training and validation were performed with KID embryos only. This methodology differs from previous investigations[46]which included discarded embryos in the KID-negative group, which constitutes inevitably an inaccurate generalization, since we cannot be sure that all discarded embryos would fail to implant.

However, some limitations of our study should be recognized. A first limitation concerns the relatively limited sample size of the study. Our approach, while adapted to narrower datasets without the risk of overfitting, would probably benefit from the inclusion of a larger number of samples. We hypothesize that sample size differences between the two validation datasets could explain, at least in part, the performances of the models (first validation with knowledge of matched embryo versus second validation without knowledge of previous transfer outcome). The difference in complexity between the two validation tasks could be another explanation for the different performances that we observed. Validation of a prediction model with a local dataset represents another limit and external validation in independent cohorts could help confirm the generalizability of the model[56]. The reliability of this model is also limited by the retrospective design of the study. Further prospective randomized studies would be needed to confirm the clinical relevance of the model. Given that implantation is only an approximate approach to live birth, these studies should ideally choose live birth as their ultimate outcome. In the limitations of the study, there could also be a potential bias in the validation cohort. Implantation rates differed between the two validation tasks; however, no direct comparison was made between these groups. Implantation was assessed only within each homogeneous cohort—one consisting of paired embryos, and the other including a single embryo per cycle. Finally, another limitation concerns the transfer protocol (fresh or frozen), because we assumed that transfer outcomes would have been the same, regardless of whether embryos were transferred fresh or frozen, which is not necessarily the case, as the outcome of a transfer may depend on the endometrial preparation[59].

In conclusion, we highlighted the potential benefits of a ML model to predict implantation in a cohort of good-quality transferred embryos with a reliable predictive performance. The ML model described in this study provides additional information that could improve the efficiency, objectivity and consistency of the embryo selection process. In this perspective, further external validation with larger datasets from other centers is needed. We believe that extending the datasets could improve the predictive performance of the model, which could be used in the near future as a clinical decision-making tool.

## Data availability

Data availability statement: Anonymized data will be made available on request to the corresponding author.

Received: 15 April 2025; Accepted: 3 July 2025
Published online: 01 August 2025

## References

1. ICMART. *'At Least 12 Million Babies' since the First IVF Birth in 1978*. (2023). https://www.focusonreproduction.eu/article/ESHRE-News-COP23_adamson
2. Fishel, S. et al. Evolution of embryo selection for IVF from subjective morphology assessment to objective time-lapse algorithms improves chance of live birth. *Reprod. Biomed. Online*. **40**, 61–70 (2020).
3. Motato, Y. et al. Morphokinetic analysis and embryonic prediction for blastocyst formation through an integrated time-lapse system. *Fertil. Steril.* **105**, 376–384e9 (2016).
4. Petersen, B. M., Boel, M., Montag, M. & Gardner, D. K. Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on day 3. *Hum. Reprod.* **31**, 2231–2244 (2016).
5. Armstrong, S. et al. Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database Syst. Rev.* **5**, CD011320 (2019).
6. Payá, E., Bori, L., Colomer, A., Meseguer, M. & Naranjo, V. Automatic characterization of human embryos at day 4 post-insemination from time-lapse imaging using supervised contrastive learning and inductive transfer learning techniques. *Comput. Methods Programs Biomed.* **221**, 106895 (2022).
7. Berntsen, J., Rimestad, J., Lassen, J. T., Tran, D. & Kragh, M. F. Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences. *PLoS One*. **17**, e0262661 (2022).

8. Bori, L. et al. Novel and conventional embryo parameters as input data for artificial neural networks: an artificial intelligence model applied for prediction of the implantation potential. *Fertil. Steril.* **114**, 1232–1241 (2020).
9. Tran, D., Cooke, S., Illingworth, P. J. & Gardner, D. K. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Hum. Reprod.* **34**, 1011–1018 (2019).
10. Ahlström, A. et al. Correlations between a deep learning-based algorithm for embryo evaluation with cleavage-stage cell numbers and fragmentation. *Reprod. Biomed. Online*. **47**, 103408 (2023).
11. Liu, H. et al. Development and evaluation of a live birth prediction model for evaluating human blastocysts from a retrospective study. *eLife* 12, e83662 (2023).
12. Ueno, S., Berntsen, J., Ito, M., Okimura, T. & Kato, K. Correlation between an annotation-free embryo scoring system based on deep learning and live birth/neonatal outcomes after single vitrified-warmed blastocyst transfer: a single-centre, large-cohort retrospective study. *J. Assist. Reprod. Genet.* **39**, 2089–2099 (2022).
13. Wang, G. et al. A generalized AI system for human embryo selection covering the entire IVF cycle via multi-modal contrastive learning. *Patterns (N Y)*. **5**, 100985 (2024).
14. Boyer, P. & Boyer, M. [Non invasive evaluation of the embryo: morphology of preimplantation embryos]. *Gynecol. Obstet. Fertil.* **37**, 908–916 (2009).
15. Gardner, D. K. & Schoolcraft, W. B. In Vitro culture of human blastocyst. in *Towards Reproductive Certainty: Infertility Genet. Beyond* 377–388 (1999).
16. Ciray, H. N. et al. Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. *Hum. Reprod.* **29**, 2650–2660 (2014).
17. Bori, L. et al. The higher the score, the better the clinical outcome: retrospective evaluation of automatic embryo grading as a support tool for embryo selection in IVF laboratories. *Hum. Reprod.* **37**, 1148–1160 (2022).
18. Ferreux, L. et al. Live birth rate following frozen-thawed blastocyst transfer is higher with blastocysts expanded on day 5 than on day 6. *Hum. Reprod.* **33**, 390–398 (2018).
19. Corroenne, R. et al. Endometrial Preparation for frozen-thawed embryo transfer in an artificial cycle: transdermal versus vaginal Estrogen. *Sci. Rep.* **10**, 985 (2020).
20. Li, C. et al. YOLOv6: A Single-Stage object detection framework for industrial applications. (2022). https://doi.org/10.48550/ARXIV.2209.02976
21. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *Proc. 37th Int. Conf. Mach. Learn. (ICML'20)*. **119**, 1597–1607 (2020).
22. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations (ICLR)* 1–14 (2015) ) 1–14 (2015) (2015). https://doi.org/10.48550/ARXIV.1409.1556
23. Gomez, T. et al. A time-lapse embryo dataset for morphokinetic parameter prediction. *Data Brief.* **42**, 108258 (2022).
24. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, Las Vegas, NV, USA, 2016). 770–778 (IEEE, Las Vegas, NV, USA, 2016). (2016). https://doi.org/10.1109/CVPR.2016.90
25. Trifonova, O. P., Lokhov, P. G. & Archakov, A. I. Metabolic profiling of human blood. *Biochem. Mosc. Suppl. Ser. B*. **7**, 179–186 (2013).
26. Barrie, A. et al. Examining the efficacy of six published time-lapse imaging embryo selection algorithms to predict implantation to demonstrate the need for the development of specific, in-house morphokinetic selection algorithms. *Fertil. Steril.* **107**, 613–621 (2017).
27. Diakiw, S. M. et al. An artificial intelligence model correlated with morphological and genetic features of blastocyst quality improves ranking of viable embryos. *Reprod. Biomed. Online*. **45**, 1105–1117 (2022).
28. Illingworth, P. J. et al. Deep learning versus manual morphology-based embryo selection in IVF: a randomized, double-blind noninferiority trial. *Nat. Med.* https://doi.org/10.1038/s41591-024-03166-5 (2024).
29. Khosravi, P. et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *Npj Digit. Med.* **2**, 21 (2019).
30. Enatsu, N. et al. A novel system based on artificial intelligence for predicting blastocyst viability and visualizing the explanation. *Reproductive Med. Biology*. **21**, e12443 (2022).
31. Ueno, S. et al. Pregnancy prediction performance of an annotation-free embryo scoring system on the basis of deep learning after single vitrified-warmed blastocyst transfer: a single-center large cohort retrospective study. *Fertil. Steril.* **116**, 1172–1180 (2021).
32. Chavez-Badiola, A., Flores-Saiffe-Farías, A., Mendizabal-Ruiz, G., Drakeley, A. J. & Cohen, J. Embryo ranking intelligent classification algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. *Reprod. Biomed. Online*. **41**, 585–593 (2020).
33. Yang, H. Y. et al. BlastAssist: a deep learning pipeline to measure interpretable features of human embryos. *Hum. Reprod.* **39**, 698–708 (2024).
34. Wang, S., Chen, L. & Sun, H. Interpretable artificial intelligence-assisted embryo selection improved single-blastocyst transfer outcomes: a prospective cohort study. *Reprod. Biomed. Online*. **47**, 103371 (2023).
35. Zhao, M. et al. Application of convolutional neural network on early human embryo segmentation during in vitro fertilization. *J. Cell. Mol. Med.* **25**, 2633–2644 (2021).
36. Zhao, M. et al. Automated and precise recognition of human zygote cytoplasm: A robust image-segmentation system based on a convolutional neural network. *Biomed. Signal Process. Control*. **67**, 102551 (2021).
37. Ezoe, K. et al. Association between a deep learning-based scoring system with morphokinetics and morphological alterations in human embryos. *Reprod. Biomed. Online*. **45**, 1124–1132 (2022).
38. Kato, K. et al. Does embryo categorization by existing artificial intelligence, morphokinetic or morphological embryo selection models correlate with blastocyst euploidy rates? *Reprod. Biomed. Online*. **46**, 274–281 (2023).
39. Bamford, T. et al. A comparison of 12 machine learning models developed to predict ploidy, using a morphokinetic meta-dataset of 8147 embryos. *Hum. Reprod.* **38**, 569–581 (2023).
40. Zou, Y. et al. Can the combination of time-lapse parameters and clinical features predict embryonic ploidy status or implantation? *Reprod. Biomed. Online*. **45**, 643–651 (2022).
41. Blank, C. et al. Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. *Fertil. Steril.* **111**, 318–326 (2019).
42. Sawada, Y. et al. Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth. *Reprod. Biomed. Online*. **43**, 843–852 (2021).
43. Miyagi, Y., Habara, T., Hirata, R. & Hayashi, N. Feasibility of predicting live birth by combining conventional embryo evaluation with artificial intelligence applied to a blastocyst image in patients classified by age. *Reprod. Med. Biol.* **18**, 344–356 (2019).
44. Duval, A. et al. A hybrid artificial intelligence model leverages multi-centric clinical data to improve fetal heart rate pregnancy prediction across time-lapse systems. *Hum. Reprod.* **38**, 596–608 (2023).
45. Ueno, S., Berntsen, J., Okimura, T. & Kato, K. Improved pregnancy prediction performance in an updated deep-learning embryo selection model: a retrospective independent validation study. *Reprod. Biomed. Online*. **48**, 103308 (2024).
46. Theilgaard Lassen, J., Fly Kragh, M., Rimestad, J., Nygård Johansen, M. & Berntsen, J. Development and validation of deep learning based embryo selection across multiple days of transfer. *Sci. Rep.* **13**, 4235 (2023).

47. Fruchter-Goldmeier, Y. et al. An artificial intelligence algorithm for automated blastocyst morphometric parameters demonstrates a positive association with implantation potential. *Sci. Rep.* **13**, 14617 (2023).
48. Erlich, I. et al. Pseudo contrastive labeling for predicting IVF embryo developmental potential. *Sci. Rep.* **12**, 2488 (2022).
49. Cohen, J. et al. Artificial intelligence in assisted reproductive technology: separating the dream from reality. *Reprod. Biomed. Online*. **50**, 104855 (2025).
50. Salih, M. et al. Embryo selection through artificial intelligence versus embryologists: a systematic review. *Hum Reprod Open* hoad031 (2023). (2023).
51. Glatstein, I., Chavez-Badiola, A. & Curchoe, C. L. New frontiers in embryo selection. *J. Assist. Reprod. Genet.* **40**, 223–234 (2023).
52. Lee, T., Natalwala, J., Chapple, V. & Liu, Y. A brief history of artificial intelligence embryo selection: from black-box to glass-box. *Hum. Reprod.* **39**, 285–292 (2024).
53. Kragh, M. F. & Karstoft, H. Embryo selection with artificial intelligence: how to evaluate and compare methods? *J. Assist. Reprod. Genet.* **38**, 1675–1689 (2021).
54. Dimitriadis, I., Zaninovic, N., Badiola, A. C. & Bormann, C. L. Artificial intelligence in the embryology laboratory: a review. *Reprod. Biomed. Online*. **44**, 435–448 (2022).
55. Bori, L. et al. An artificial intelligence model based on the proteomic profile of euploid embryos and blastocyst morphology: a preliminary study. *Reprod. Biomed. Online*. **42**, 340–350 (2021).
56. Zhu, J. et al. External validation of a model for selecting day 3 embryos for transfer based upon deep learning and time-lapse imaging. *Reprod. Biomed. Online*. **47**, 103242 (2023).
57. Grinsztajn, L., Oyallon, E. & Varoquaux, G. *Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?* (2022). https://doi.org/10.48550/arXiv.2207.08815
58. Shwartz-Ziv, R. & Armon, A. Tabular data: deep learning is not all you need. *Inform. Fusion*. **81**, 84–90 (2021).
59. Roque, M., Haahr, T., Geber, S., Esteves, S. C. & Humaidan, P. Fresh versus elective frozen embryo transfer in IVF/ICSI cycles: a systematic review and meta-analysis of reproductive outcomes. *Hum. Reprod. Update*. **25**, 2–14 (2019).


## Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011011303R4 made by GENCI. We are grateful to Jenessa Pettit for editing the English version of the manuscript.

## Author contributions

Conceptualization: P.M.P., L.B. and F.C.; Methodology: F.C., L.B. and P.M.P.; Software: F.C.; Validation: P.E.B. and P.R.; Investigation: L.B., F.C. and M.B.; Writing - original draft preparation: L.B. and F.C.; Writing - review and editing: F.C., L.B., M.B. and P.M.P.; Supervision: P.M.P., P.R. and P.E.B. All authors have read and approved the published version of the manuscript.

## Funding

No external funding was received for conducting this study.

## Declarations

## Competing interests

The authors declare no competing interests.

## Informed consent

Informed consent was obtained from participants before inclusion. Patients who objected to the processing of their data for research were excluded. The study protocol was reviewed and approved by the Ethics Committee of Angers University Hospital, France (ref. no. 2024-045, reviewed on March 14, 2024). The anonymized database was registered in the CNIL register of the Univeristy Hospital of Angers (ar24-0069v0) in compliance with the General Data Protection Regulation (GDPR). All procedures and protocols complied with France regulations and with the Helsinki Declaration guidelines.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-10531-y.

**Correspondence** and requests for materials should be addressed to L.B.