

BEN-GURION UNIVERSITY OF THE NEGEV
FACULTY OF ENGINEERING SCIENCES
DEPARTMENT OF SOFTWARE AND INFORMATION ENGINEERING

Using unlabeled information of embryo siblings from the same cohort cycle to enhance in
vitro fertilization implantation prediction

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
M.Sc DEGREE

By: Noam Tzukerman

September 2022

BEN-GURION UNIVERSITY OF THE NEGEV

FACULTY OF ENGINEERING SCIENCES
DEPARTMENT OF SOFTWARE AND INFORMATION ENGINEERING

Using unlabeled information of embryo siblings from the same cohort cycle to enhance in vitro fertilization implantation prediction

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
M.Sc DEGREE

By: Noam Tzukerman

Supervised by: Dr. Assaf Zaritsky

Author: *NT*.....

Date: 7/11/2022.....

Supervisor: *AZ*.....

Date: 7/11/2022.....

Chairman of the Graduate Studies Committee:.....

Date:.....

September 2022

Abstract

High content time-lapse embryo imaging assessed by machine learning is revolutionizing the field of in vitro fertilization (IVF) transfers. However, the vast majority of IVF embryos are not transferred to the uterus, and these masses of embryos with unknown implantation outcomes are ignored in current efforts that aim to predict implantation. Here, we explore whether, and to what extent the information encoded within “sibling” embryos from the same IVF cohort contribute to the performance of machine learning-based implantation prediction. First, we show that the implantation outcome is correlated with attributes derived from the siblings cohort. Second, we demonstrate that this unlabeled data boosts implantation prediction performance. Third, we characterize the siblings cohort properties driving embryo prediction, especially those that rescued erroneous predictions. Our results suggest that predictive models for embryo implantation can benefit from the overlooked, widely available unlabeled data of sibling embryos by reducing inherent noise of the individual transferred embryo.

Keywords

IVF cycle, cohort siblings, Gardner grading system, phenotypic correlations, time-lapse video microscopy, unlabeled data, embryo images, implantation prediction, machine learning-based embryo selection, morphological annotations, morphokinetic annotations, blastocyst, (AI)-based embryo assessment, image processing, Wilcoxon signed-rank test, Mann-Whitney-U signed rank test, siblings similarity.

Acknowledgements

First, I would like to express my deep and sincere gratitude to my adviser, Dr. Assaf Zaritsky, for the continuous support and guidance through the whole process, and for giving me the opportunity to conduct research and providing me with invaluable guidance.

I would also like to thank AiVF Ltd research division and particularly Maya Tsarfati Shapiro, Ron Maor, Daniella Gilboa, Daniel S. Seidman and Marcos Meseguer for the successful collaboration with Zaritsky Lab, and for their inputs and suggestions in every stage of the process.

Finally, I would like to thank Nadav Rappoport, Zeev Bomzon, and Yishay Tauber for critically reading the manuscript.

Contents

List of Figures	11
Introduction	12
Previous Work	14
Clinical background	14
EmbryoScope	14
The blastocyst	14
Related work	17
IVF procedure overview	17
Embryo selection	18
Morphology-based AI for embryo selection	18
Morphokinetic-based AI for embryo selection	19
Deep learning based AI for embryo selection	19
Contribution of sibling information to embryo implantation prediction	21
Results	23
Embryos from the same cohort are phenotypically correlated	23
Cohort properties correlate with implantation outcome	25
Cohort properties contribute to implantation prediction	27
Identifying cohort properties driving the model's prediction	29
Identifying cohort properties that corrected erroneous prediction	31
Discussion	35
Methods	38
Experiments	38
Data collection and ethics	38
Annotation of embryo clinical quality	38
Analysis	39
Embryo morphological feature representation	39
Embryo morphokinetic feature representation	39
Similarity between timing of morphokinetic events	40
Cohort morphology-based feature representation	40

Automated deep learning based embryo implantation prediction	41
Images Preprocessing	41
Image classification model	41
Evaluation of machine learning models without versus with cohort features	42
Feature importance analysis	42
References	44
Supplementary Tables	51
Supplementary Figures	52

Terminology

Term	Definition
IVF	In vitro fertilization (IVF) is the joining of a woman's egg and a man's sperm in a laboratory dish.
Oocyte	A female egg cell involved in reproduction.
Oocyte retrieval	A procedure in which mature eggs are collected from the female's ovaries.
Cohort	Oocytes retrieved in the same IVF cycle.
AUC	Area Under the Curve. A way to measure a classification model's performance.
Blastocyst	A structure formed in the early development of an embryo. In humans, blastocyst formation begins about 5 days after fertilization.
Morula	An early-stage embryo consisting of 16 cells. In humans, the morula stage begins about 3-4 days after fertilization.
Blastula	A hollow sphere of cells surrounding an inner fluid-filled cavity.
Cleavage	Division of cells in the early embryo, following fertilization. Cleavage ends with the formation of the blastula.
Blastulation	Hollow sphere of cells, that produces the blastula.
Embryo culture	Embryo culture is the sterile isolation and growth of an immature or mature embryo in vitro.
Embryo transfer	The process of returning an embryo back into the uterus with the intention of achieving a successful pregnancy.
CNN	Convolutional neural network. A class of artificial neural networks, most applied to analyze visual imagery.

Time-lapse microscopy system (TLS)	A system that can take digital images of embryos at frequent time intervals.
EmbryoScope	A time-lapse incubator that has a built-in camera and microscope. It captures images of the developing embryos every 10-15 minutes and creates a time-lapse video of embryo development.
Fertilization	A sperm fusing with an egg cell.
Hidden layers	Layers in between input layers and output layers.
Auto-encoder	A type of artificial neural network used to learn a lower dimensional representation (encoding) for a higher dimensional data, typically for dimensionality reduction.
Implantation	Implantation is when a fertilized egg, or blastocyst, is attached to the lining of the uterine wall. It marks the beginning of pregnancy.
Inner cell mass - ICM	The mass of cells inside the embryo that will eventually give rise to the definitive structures of the fetus.
Trophectoderm	A layer of cells on the outer edge of a blastocyst which develop into a large part of the placenta. These cells provide nutrients for the developing embryo.
Morphokinetics	Time specific morphological changes during embryo development. Provide dynamic information on a fertilized egg.
Oocyte age	Age of the female from which the oocytes were retrieved.
Zygote	Fertilized egg cell that results from the union of a female egg with a male sperm. The zygote stage is brief and is followed by cleavage, when the single cell becomes subdivided into smaller cells.
Pronuclei	A pronuclei is a structure representing the sperm and egg cell DNA during the process of fertilization. The sperm cell becomes a pronucleus after the sperm enters the ovum, but before the

	genetic material of the sperm and egg fuse. The appearance of two pronuclei is usually observed 16-18 hours after fertilization..
Random forest	An ensemble learning method for classification, regression and other tasks that operate by constructing multiple decision trees.
XGBoost	XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.
ANN	Artificial neural networks. An ANN is based on a collection of connected units or nodes called artificial neurons.

List of Figures

Figure 1: Examples of blastocyst grading.	15
Figure 2: Sibling embryos from the same cohort are more similar than non-siblings in terms of their differences in time intervals or normalized distance between consecutive morphokinetic events.	24
Figure 3: Siblings in positive cohorts are of higher morphological quality than those in negative cohorts.	26
Figure 4: Implantation prediction performance (AUC) comparison of different model pairs with versus without siblings' features.	28
Figure 5: Model explainability analysis.	31
Figure 6: Analysis of cohort properties that “rescued” erroneous prediction using a model trained with morphology, morphokinetics and oocyte age without and with cohort features.	33
Figure S1: Sibling embryos from the same cohort are more similar than non-siblings in terms of their differences in time intervals or normalized distance between fertilization and each morphokinetic event.	52
Figure S2: Analysis of cohort properties that “rescued” erroneous prediction using a model trained with deep convolutional neural network, morphology, morphokinetics and oocyte age without and with cohort features.	53
Figure S3: Distribution of oocyte age for all treatments.	54
Figure S5: Oocyte segmentation pipeline.	54

Introduction

Phenotypic variation is inherent in every biological system. A phenotype is determined by a combination of genetic and environmental factors. For example, the proportion of shared genetic background between human siblings can explain most of the variability in height [1], while environmental factors can affect gut microbiota composition [2], gene expression and disease susceptibility [3], and even lead to phenotypic variation in monozygotic (aka "identical") twins [4]. Thus, confinement of the genetic and environmental variability, for example, by considering siblings raised under similar conditions, can lead to reduced phenotypic variability, i.e., increased phenotypic similarity. We hypothesized that such phenotypic correlations can provide predictive value regarding an individual's future phenotypic state by considering the phenotypic states of its siblings.

Specifically, during in vitro fertilization (IVF) a cohort of "sibling" oocytes, all sharing the same "parents" within the same IVF treatment, are fertilized and incubated for up to six days under the same laboratory conditions before one or a few embryos from the cohort are selected for transfer into the uterus. IVF embryo phenotypes are heavily affected by genetic [5] and environmental [6] factors, and thus the genetic and environmental variability is minimized for siblings from the same cohort.

Recent advances in time-lapse video microscopy for live embryo imaging has transformed IVF into a data-intensive field. This has led to innovative attempts to automatically and unbiasedly estimate the implantation potential of embryos based on algorithmic assessment of their visual phenotypes and/or developmental trajectory [7,8,9,10,11,12]. Specifically, supervised machine learning has emerged as a powerful approach where features, computationally extracted from embryo images with known implantation outcomes, are used to train computational models to predict implantation [13,14,15,16]. These models reach performance comparable and even exceeding those of embryologists [17,18,19].

However, from the available cohort, only one or at most two embryos are selected for transfer to the uterus. This poses a major limitation for machine learning-based embryo selection approaches because the implantation potential of the majority of deselected embryos remains unknown [16]. Thus, the number of embryos with known implantation outcomes available for training is severely limited. Recent studies have attempted to overcome this limitation by

including “unlabeled” embryos in their model training schemes, specifically focusing on the subgroup of embryos found unsuitable for transfer to the uterus due to their poor morphological appearance and presumed limited implantation potential [20,21]. Other studies used the morphological annotations of embryos found unsuitable for transfer to the uterus to train models to predict the value of these morphological measurements as a readout for successful implantation potential [22,23]. However, none of these studies fully capitalized on and systematically assessed the potential of using the association between “sibling” embryos from the same cohort to enhance implantation prediction accuracy.

We asked whether we can take advantage of cohort-derived information to train a machine learning model tasked with predicting implantation? We hypothesized that cohort embryos share information relevant for implantation prediction. Thus, using the full extent of the available unlabeled data in the cohort may provide further statistical discriminative power. Indeed, a few earlier studies provide evidence supporting the notion that cohort siblings encapsulate information that correlate with the transferred embryo’s quality and implantation outcome, such as cohort size [24,25,26,27], sibling blastocyst development [28,29], or a combination of cohort-specific variables [26]. Here, we explicitly assess the contribution of sibling information to embryo implantation prediction by systematically evaluating different models trained with or without information from the cohort. We demonstrate that the unlabeled cohort embryos contribute to the prediction. Our results imply that artificial intelligence (AI)-based embryo assessment can benefit from the widely available, and currently ignored, correlated data in the cohort’s siblings. We characterize the specific properties of the cohort that contribute to the prediction and show that different cohort features can be exploited to enhance performance of different models in varying contexts. Altogether, we suggest that considering the correlated properties of sibling embryos aid in canceling the inherently noisy single embryo prediction.

Previous Work

Clinical background

EmbryoScope

Embryos images were acquired using the EmbryoScope® time-lapse system. EmbryoScope is an incubator that allows embryologists to observe the embryos development in vitro while mimicking the uterus's physiological conditions. It has a built-in camera that takes pictures of the embryos at fixed time intervals. Those pictures are then combined to a continuous video stream of the embryo development, enabling to determine and systematically characterize the exact times of important embryo developmental stages.

The blastocyst

The term blastocyst refers to a development stage that a human embryo reaches at day 5 or 6 after fertilization. It is a structure formed of hollow cells surrounding a mass of cells known as the inner cell mass (ICM) (Fig. 1). Studies have shown implantation rate is higher in blastocysts transfer compared to early-stage embryos [30,31,32]. Blastocyst transfer also enables better selection of the best quality embryo from a cohort, providing more confidence compared to the high early uncertainty in earlier developmental stages, thus reducing the need to transfer multiple embryos to achieve pregnancy, and avoiding multiple gestations. The need for morphological assessment of blastocysts for transfer contributed to the development of blastocysts scoring systems. The most popular grading system adopted by most IVF laboratories is the grading system introduced by Gardner and Schoolcraft in 1999 [8,33]. Morphological embryo assessments in this thesis were also based on this system.

According to this method the embryo is classified based on three separate quality scores:

- i) Stage of blastocyst development, expansion and hatching, ranking from 1-6, 1 being the embryo that hasn't reached the blastocyst stage yet and 6 the blastocyst hatched out of its shell.

- ii) Inner cell mass (ICM) quality, size and compactness, ranking from A to D, D being very few cells and A the embryo has many cells and is slightly packed.
- iii) Quality, cohesiveness and number of trophoctoderm (TE), ranking from A to D, D being very few large cells and A the embryo has many cells and forming a cohesive layer

Examples of blastocysts graded according to the Gardner & Schoolcraft grading system are shown in Fig. 1.

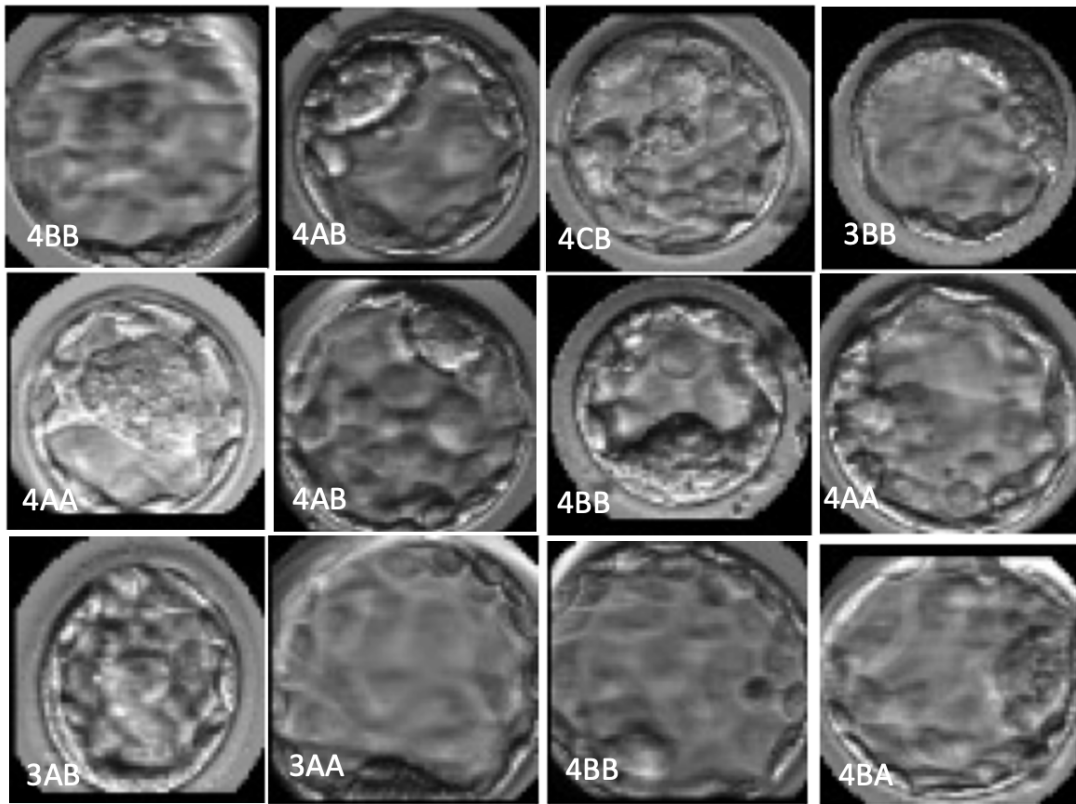


Figure 1. Examples of blastocyst grading. Blastocysts were scored and graded according to the Gardner and Schoolcraft standard [8]. Blastocysts are given scores when the expansion grade is listed first, the inner cell mass grade listed second and the trophoctoderm grade third. The Gardner grading system assigns three quality measures to each blastocyst embryo: 1. Blastocyst development, expansion, and hatching stage. 2. Inner cell mass (ICM) quality. 3. Trophoctoderm (TE) quality.

Icm grade	Inner cell mass quality
A	Many cells, tightly packed
B	Several cells, loosely grouped
C	Very few cells
D	Degenerative

Table 1: ICM score, Gardner and Schoolcraft, 1999 [8]. ICM cells are the cells located inside the blastocoel, often forming a cell clump at one pole of the blastocyst.

Trophectoderm grade	Trophectoderm quality
A	Many cells, forming a cohesive layer
B	Few cells, forming a loose epithelium
C	Very few large cells
D	Degenerative

Table 2: Trophectoderm score, Gardner and Schoolcraft, 1999 [8]. Trophectoderm cells are the outer cells of the blastocyst, forming the blastocyst structure itself.

Expansion degree	Development and expansion stage
BT-early blastocyst	Blastocoel cavity more than half the volume of the embryo
BC -full blastocyst	Full blastocyst, cavity completely filling the embryo

BE-expanded blastocyst	Expanded blastocyst, cavity larger than the embryo, with thinning of the
Bhi-hatching blastocyst	Hatching out of the shell
BH-hatched blastocyst	Hatched out of the shell

Table 3: Development and expansion stage, Gardner and Schoolcraft, 1999 [8].

Related work

IVF procedure overview

In vitro fertilization (IVF) is a procedure used for infertility treatment, gestational surrogacy or for preventing genetic problems. This process involves retrieving eggs from a woman’s ovaries and fertilizing them with sperm. The fertilized egg, the zygote, is then monitored for 2–6 days in an in vitro laboratory culture environment. After 3-6 days it is transferred back into the uterus, with the intention of achieving a successful pregnancy. Usually, IVF embryos are maintained in vitro until cleavage stage (day 3 after fertilization) or blastocyst stage (day 5 or 6 after fertilization).

Although IVF is the most effective method of assisted reproductive technology, only one-third of all IVF cycles result in a pregnancy [34]. Besides the fact that it may take more than one cycle for a successful pregnancy, it is also an expensive and a highly emotive and stressful experience for patients. All those reasons emphasize the need for an improved and accurate embryo selection process. The number of embryos to be transferred is determined by the patient age, quality of the embryos and other diagnostic factors. Transferring more than one embryo may result in a multiple gestation (multiple pregnancy). Multiple gestation is associated with maternal and embryo complications, especially preterm delivery that increases risk of significant neonatal morbidity and mortality [35,36]. Today, most clinics try to minimize the number of transferred embryos to avoid it.

Embryo selection

Embryo selection made by human embryologists can suffer from human bias in critical decisions that would affect the success of the IVF cycle. The subjectivity involved in the embryos evaluation process, leads to results that vary across clinics and embryologists. AI methods are used more frequently in recent years for improving human embryo selection and assessing the probability of an embryo to be implanted back in the uterus in in-vitro fertilization treatments [34]. These are based on two sets of visual embryo features: morphology, the visual appearance of an embryo, and morphokinetics, the timing of a key embryo developmental stage.

Morphology-based AI for embryo selection

Uyar et al. (2015) [17] used the Naïve Bayes classifier for predicting embryos implantation probabilities using morphological features, patient demographics and treatment characteristics, as infertility factor and treatment protocol. They used an information gain feature weighting approach for reducing the number of features. The model results were compared with the predictions of five embryologists' majority decision regarding the implantation outcome. The model outperformed the expert's judgment in terms of accuracy and false alarm rate. Hassan et al. (2020) [37] assessed the implantation prediction of five different machine learning models: Multilayer Perceptron (MLP), Support Vector Machines (SVM), C4.5, Classification and Regression Trees (CART) and Random Forest (RF). According to their experiments, SVM attained the highest accuracy.

Another study [38] compared the performance of five different models: Support vector machines (SVM), recursive partitioning (RPART), random forest (RF), Adaptive boosting (Adaboost), and 1NN and found that in this case SVM also outperformed the other models. Guh et al. (2011) [39] used Genetic Algorithms to choose the optimal subset of clinical attributes and the corresponding hyper-parameters for a decision tree model that predicts IVF cycle outcome.

Blank et al. (2019) [27] compared both random forest and multivariate logistic regression models to predict implantation of a day 5 blastocyst. They used embryos morphologic characteristics and

patients characteristics for this purpose. Random forest achieved a much better AUC score compared to the multivariate logistic regression model.

Morphokinetic-based AI for embryo selection

The recent development of the EmbryoScope has led to prediction methods based on morphokinetic parameters shown to significantly increase ongoing pregnancy and implantation rates [7,40]. Specifically, the timing of blastulation has been shown to be an important predictor for implantation [33,41,42]. Increased time intervals were also positively associated with poor embryo quality, reduced blastomere number and increased fragmentation [10]. Morphokinetic events are usually referred to as tN, which is the time of cell division to the N-cell stage. Wu et al. (2016) [43] characterized morphokinetic parameters of transferred cleavage embryos based on their implantation outcome. Results showed that tPNf (the time when both pronuclei disappear), t2 and t4 occurred at an earlier stage in embryos with successful implantation. In addition, the time interval between t3 and t4 was significantly shorter for successfully implanted embryos. Meseguer et al. (2011) [7] examined transferred cleavage for the purpose of identifying the most promising variables characterizing implanted embryos. They reported that t5, t3-t2 and t4-t3 were the most predictive parameters. t2, t3, t4, t5 and t3-t2 were normally distributed for implanted embryos, as opposed to non-implanted embryos. They have divided the embryos into subcategories twice: the first was based on morphology and the second was based on morphokinetic. Even though in both divisions the best subcategory had also the highest implantation rate, the morphokinetic's best subcategory had higher implantation rate than the best morphologic one.

Deep learning based AI for embryo selection

Artificial Neural Networks (ANN) were trained with both morphokinetic and morphological parameters to predict live birth and implantation potential [44,45]. Durairaj and Thamilselvan (2013) [24] used ANN for implantation prediction with parameters based on the women's

physiology and treatments history, such as number of oocytes retrieved, number of embryos transferred and previous history of miscarriages.

The development of the EmbryoScope has enabled the evolution of image processing techniques on time-lapse imaging for embryo selection and assessment. Tran et al. (2019) [16] studied the use of IVY, a deep-learning model that can analyze whole time-lapse videos instead of specific images of blastocysts, to predict implantation probability. They trained the model to discriminate between successfully implanted embryos versus embryos which failed to implant or have been discarded. Bormann et al. (2020) [46] have investigated the use of a CNN pre-trained with ImageNet images and used transfer learning for blastocysts single time-point images. In the first stage, they classified embryos into categories based on the Gardner grading system using a pre-trained network [47] and a genetic algorithm. Another recently developed framework is UBar, a predictive model for embryos implantation potential [18]. UBar uses a CNN auto-encoder to derive a 968-dimensional latent vector of developing embryos in a sequence of time-lapse frames. A LSTM network was then provided with the embeddings data. The network was trained on 4087 embryos, which had a known embryologist's grade, but not all of them had a known implantation outcome. Then, the network was used to predict implantation of 272 embryos. Results showed UBar has outperformed a panel of five embryologist's decisions. Khosravi et al. (2019) [19] implemented STORK, a robust framework containing a DNN network and transfer learning architecture which was trained on a large dataset embryos time-lapse image, for predicting blastocyst quality and selecting the best embryo for transfer. They classified 10,148 embryos into three quality groups: poor, fair, and good. STORK was evaluated on 1930 images of 283 embryos. Although it was not designed to predict implantation, it could detect embryo quality based on morphological classification and associate it with pregnancy rate. STORK outperformed a majority voting procedure of 5 embryologists in terms of predicting implantation, thus providing evidence for the potential superiority of the machine prediction over the human embryologist in this task.

Semi-supervised learning

When unlabeled data is used with a small amount of labeled data, it can produce considerable improvement in models' performance. This approach is called semi-supervised learning, and it exploits the advantages of both unsupervised and supervised learning while avoiding the challenges of finding a large amount of labeled data. An example of when semi-supervised learning can be used successfully is in the building of a text document classifier [64,65].

Semi-supervised learning is ideal in this case because it is nearly impossible to find a large amount of labeled text documents. Our work is an example of utilizing semi-supervised learning, where a small fraction of labeled (transferred embryos) and vast unlabeled observations (non-transferred cohort siblings) are used together to construct better learning procedures. In our case, the unlabelled data from the siblings is useful because it carries information useful for the transferred embryo's label prediction that is not contained in the labeled data alone.

Contribution of sibling information to embryo implantation prediction

Previous work has shown that cohort data contain information helpful for assessing the embryo's quality and evaluating its implantation competence. A few earlier studies showed that cohort size is correlated with implantation outcome [24,26,27]. Devreker et al. (1999) [25] discovered that a large embryo cohort size was the most predictive parameter for having at least one good quality embryo in the cycle, independent of oocyte age.

Researchers have tested if having a blastocyst sibling in the treatment affects the probability of an embryo to be implanted and results in a live birth. Fisch et al. (1999) [28] discovered that having at least one blastocyst in the cohort resulted in a higher number of oocytes retrieved, higher blastocyst embryos rate in the treatment and increased clinical pregnancy rates. Mackenna et al. (2013) [29] have also shown that having a blastocyst sibling is a prognostic factor for implantation. They examined patients who had embryos transferred back to the uterus at day 3 and had remaining siblings' embryos in culture. They evaluated the remaining cultured embryos at day 5 and divided the patients into two groups: patients with at least one blastocyst embryo and patients with no blastocyst embryos. The first group had significantly higher clinical pregnancy rate per cycle, live birth rate and implantation rate. No differences were observed in

multiple pregnancy rates and miscarriage rates between both groups. Jun et al. (2008) [26] claimed that four cohort parameters were remarkably more informative than any measures of individual transferred embryos in predicting an IVF cycle outcome: cohort size, number of 8-cell embryos, percentage of cleavage arrest in the cohort, and day 3 FSH (Follicle-Stimulating Hormone) level. They used Sequential Multiple Additive Regression Tree (MART) and Classification and Regression Tree (CART) models trained only with those four cohort features and reported results of 70% accuracy in predicting implantation outcome. Furthermore, they showed that most prognostic information carried by conventional parameters such as oocyte age and clinical diagnoses, is captured by three of the four parameters. However, they did not analyze siblings' information on day-5 stage, but only on day-3.

In my thesis I show that extracting information from siblings in the same cohort cycle of the transferred embryo can improve AI-based implantation prediction. Our work is different from previous studies because we evaluate the contribution of the cohort information to implantation prediction directly by training different models with and without cohort data. For each trained model, we examine the most contributed parameters and show that different cohort properties are exploited in different models. We also demonstrate that siblings are more phenotypically similar than non-sibling embryos, which none of the previous studies has researched. In addition, most of the previous studies which examined cohort properties assessed day 3 embryos. Since the uncertainty of the implantation prediction is higher at this stage of development [30,31], additional information from the siblings are expected to contribute to the implantation prediction. We are only evaluating embryos at the blastocyst stage, when the uncertainty of the implantation prediction is lower than in day 3, but we still manage to show that cohort information contributes discriminative signals beyond the transferred embryo properties. We also examine which embryos were correctly classified only by the classifier that had access to cohort information and what are their characteristics, an analysis that none of the previous studies has been conducted.

Results

Embryos from the same cohort are phenotypically correlated

Our data included information derived from 2089 transferred embryos collected from 1605 IVF cycles that included 1176 implanted blastocysts (*positive embryos*), 913 non-implanted blastocysts (*negative embryos*), and 14105 sibling embryos that were not transferred (Methods, Table S1). The timing of 7 hallmark stages in embryo development, termed morphokinetic events, were manually annotated based on time-lapse observation of the developing embryos (Fig. 2A, Table S1, Methods). These morphokinetic events are considered key in proper embryo development and were shown to be correlated with implantation potential [10,48]. The timing of morphokinetic events from fertilization and the time intervals between consecutive morphokinetic events were more similar among sibling embryos than for randomly selected non-sibling embryos, indicating lower intra-cohort variability (Fig. S1A-F - timing of morphokinetic events, Fig. 2B-F - time intervals between consecutive morphokinetic events, Fig. 2G - normalized distances between multivariate representations of all time intervals between consecutive morphokinetic events, Fig. S1G - normalized distances between multivariate representations of all morphokinetic events). To verify that this higher intra-cohort (i.e., between siblings) similarity is not an artifact due to the random selection of embryos, we compared intra-versus inter-cohort similarity in the morphokinetic profiles of embryos with similar morphological qualities. We considered the manually annotated Gardner and Schoolcraft alphanumeric quality scoring scheme [8] as a proxy for embryo morphological-based quality, which is based on the assessment of three parameters: blastocyst expansion status, morphology of the inner cell mass (ICM), and morphology of the trophectoderm (TE) [49,50] (Fig. 2H). We considered all possible ordered combinations of triplets of embryos that include a pair of sibling embryos from the same cohort and a third non-sibling embryo from a different cohort, where all three embryos had the same Gardner scores (Fig. 2I, Methods). This analysis established that morphokinetic multivariate profiles of sibling embryos are more similar than non-sibling embryos and that this similarity is not a mere consequence of a morphologic similarity between siblings at the blastocyst stage (Fig. 2J). Altogether, this data supports the notion that sibling embryos share similar phenotypic properties.

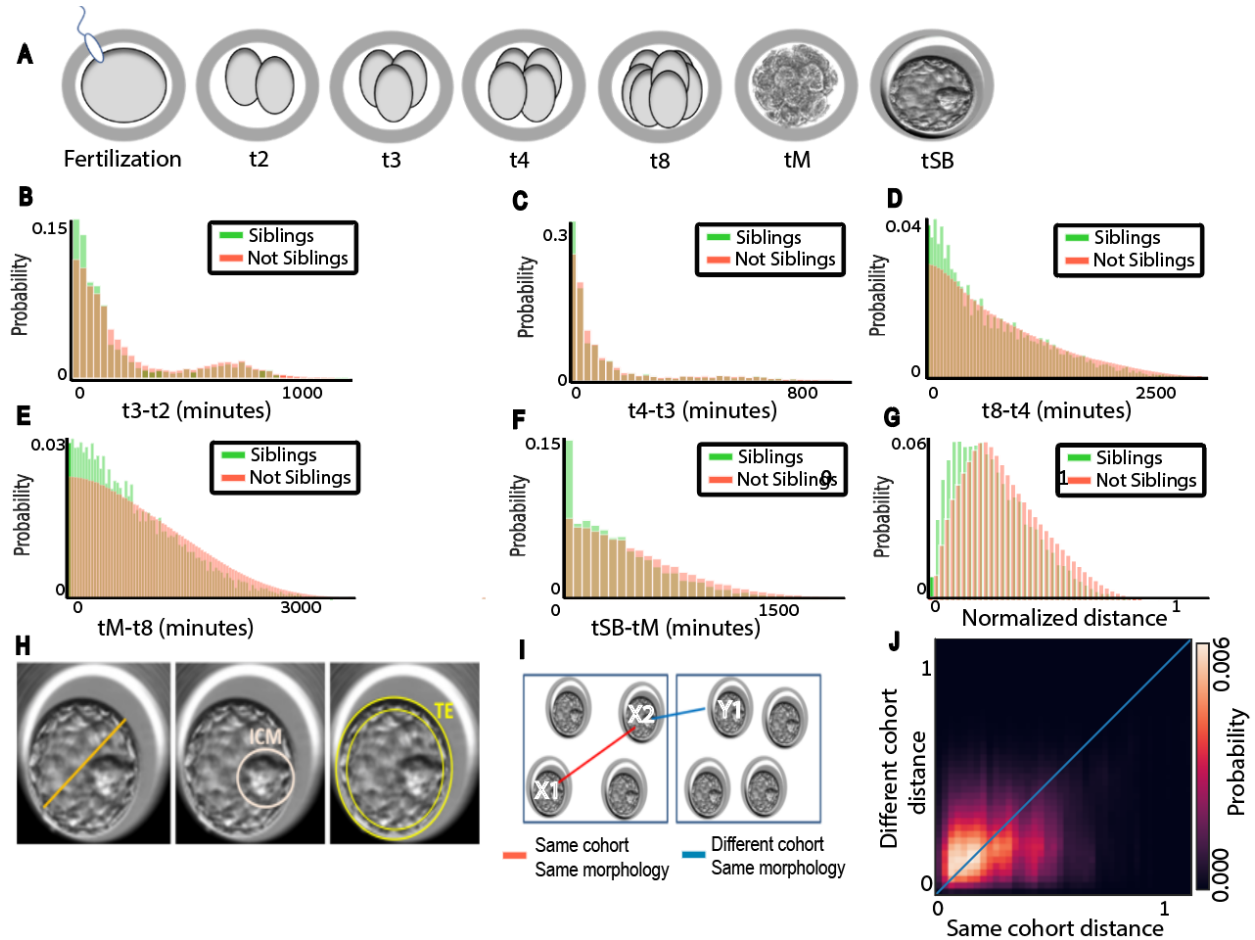


Figure 2. Sibling embryos from the same cohort are more similar than non-siblings in terms of their morphokinetic properties. (A) Schematic sketch. Morphokinetic features: cell division to the 2, 3, 4 and 8-cell stage (t2, t3, t4, t8), the compaction of the morula - a day-3 development stage (tM) and the start of blastulation (tSB) - a day-5 development stage. (B-G) Distribution of the difference in time intervals, in minutes (B-F) or normalized distance (G), between consecutive morphokinetic events compared across siblings versus not siblings embryo pairs. N embryos = 16194. N cohorts = 1605. N positive cohorts = 928, N negative cohorts = 677. (B) Mean (standard deviation) of distances between t3-t2 intervals was 174.03 (214.7) for sibling embryos versus 202.26 (218.7) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (C) Mean (standard deviation) of distances between t4-t3 intervals was 146.48 (220.56) for sibling embryos versus 158.75 (225.52) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (D) Mean (standard deviation) of distances between t8-t4 intervals was 594.58 (531.66) for sibling embryos versus 684.09 (569.68) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (E) Mean (standard deviation) of distances between tM-t8 intervals was 691.03 (546.57) for sibling embryos versus 815.16 (602.9) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (F) Mean (standard deviation) of distances between tSB-tM intervals was 344.50 (336.89) for sibling embryos versus 434.79 (372.23) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (G) Mean (standard deviation) of normalized distances between all morphokinetic features time intervals was 0.241 (0.14) for sibling embryos versus 0.28 (0.15) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (H) Predefined criteria of blastocyst quality according to the Gardner three-part scoring scheme. From left to right: Blastocyst expansion – volume and degree of expansion of the blastocyst cavity (ranked 1-6). Morphology of the Inner cell mass (ICM)– size and compaction of the mass of cells that eventually form

the fetus (ranked A-D). Morphology of the trophoctoderm (TE) – number and cohesiveness of the single cell layer on the outer edge of the blastocyst that eventually forms the placenta (ranked A-D). (I) Schematic sketch of the analysis comparing embryo triplets: two from the same cohort (annotated X1 and X2), and two from a different cohort (X2 and Y1), where all embryos have the same Gardner annotations (similar morphological quality). (J) Mean (standard deviation) of normalized distances between all morphokinetic features time intervals was 0.224 (0.141) for sibling embryos with similar Gardner scores versus 0.263 (0.146) for not-sibling embryos with similar Gardner scores, Mann-Whitney-U signed rank test p-value < 0.0001. N = 1,654,732 ordered triplets.

Cohort properties correlate with implantation outcome

The phenotypic similarity between siblings from the same cohort raises the hypothesis that morphological and morphokinetic properties of sibling embryos are correlated to the implantation outcome. To test this hypothesis we compared the distribution of several cohort-related properties for cohorts that included successfully implanted embryos (*positive cohorts*) versus those cohorts where the transferred embryo/s failed to implant (*negative cohorts*). First, we validated that positive cohorts contained more embryos than negative cohorts (Fig. 3A), and that the fraction of sibling embryos within a cohort (not including the transferred embryo/s) reaching blastulation was larger in positive cohorts (Fig. 3B). These results are in agreement with previous reports for cohort size [24,25,26,27,38] and for siblings blastocyst development [28,29]. Each of the three Gardner morphological scores was elevated for sibling embryos, in positive cohorts compared to negative cohorts (Fig. 3C-E). While the enrichments of higher quality cohort properties in positive cohorts were relatively small for each cohort property, they were all consistent toward favoring higher quality cohort properties in positive cohorts. Cumulatively, these results conclude that morphological properties of “sibling” embryos within a cohort, that were not transferred, are associated with the implantation potential of the embryo that was transferred within that cohort. These results establish that properties derived from sibling embryos correlate with the clinical outcome of their sibling transferred embryo.

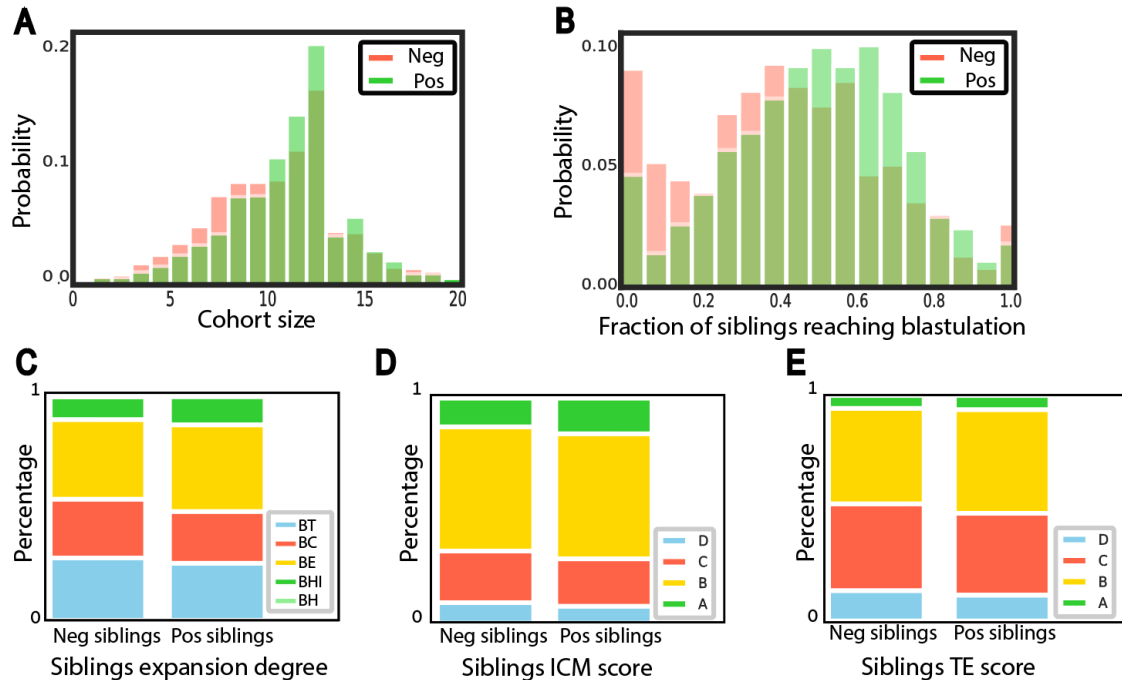


Figure 3: Siblings in positive cohorts are of higher morphological quality than those in negative cohorts. **(A-B)** Distribution of the cohort size (i.e., number of sibling embryos in a cohort) **(A)** or the fraction of embryos within a cohort (not including the transferred embryo/s) to develop into a blastocyst **(B)** compared across positive versus negative cohorts. N transferred blastocysts = 2089. N implanted blastocyst = 1176, N non-implanted blastocyst = 913. N cohorts = 1605. N positive cohorts = 928, N negative cohorts = 677. **(A)** Mean (standard deviation) cohort size was 10.56 (3.19) for positive cohorts versus 9.9 (3.36) for negative cohorts, Mann-Whitney-U signed rank test p-value < 0.0001. **(B)** Mean (standard deviation) fraction of sibling embryos within a cohort (not including the transferred embryo/s) reaching blastulation was 0.49 (0.22) for positive cohorts versus 0.41 (0.25) for negative cohorts, Mann-Whitney-U signed rank test p-value < 0.0001. **(C-E)** Distribution of manually annotated Gardner scores of cohort embryos (not including the transferred embryo/s) across positive versus negative cohorts. N transferred blastocysts = 1936 embryos. N implanted blastocysts = 1141, N non-implanted blastocysts = 795. **(C)** Expansion degree: BT- early blastocyst, BC - full blastocyst, BE- expanded blastocyst, Bhi -hatching blastocyst, BH - hatched blastocyst. The total number of BH embryos in our dataset is 3 for implanted blastocysts and 3 for non-implanted blastocysts, and thus cannot be seen in the graph. Corresponding Mann-Whitney-U signed rank tests on the null hypothesis that the two schemes were drawn from the same matched distribution p-value < 0.0001. **(D)** Inner cell mass (ICM): ranked from A (high quality) to D (low quality). Corresponding Mann-Whitney-U signed rank tests on the null hypothesis that the two schemes were drawn from the same matched distribution p-value < 0.0001. **(E)** Trophectoderm (TE): A (high quality) to D (low quality). Corresponding Mann-Whitney-U signed rank tests on the null hypothesis that the two schemes were drawn from the same matched distribution p-value < 0.0001.

Cohort properties contribute to implantation prediction

Given that cohort properties were correlated to the implantation outcome, we hypothesized that inclusion of cohort-derived features can enhance the prediction power of a machine learning model initially trained without cohort information (Fig. 4A). To assess the generality of this idea we trained several distinct machine learning models for the prediction of implantation outcome, each of these models was trained without and with cohort-derived features. The performance of each pair of models, without or with cohort features, was compared to assess the contribution of the cohort information. The first model was morphology-based, and trained on manually annotated Gardner scores (Methods). The second model was morphokinetics-based [7], and trained on manually annotated key morphokinetic events (Methods). The third model combined morphology, morphokinetics and the oocyte age, where the latter is widely accepted as correlated with implantation success [51,52] (Methods). Seventeen cohort-derived features were calculated from the siblings of the transferred embryo at test. These included cohort size, fraction of sibling embryos reaching blastulation, and features encoding the siblings' Gardner scores (Methods). The performance of each of these three models was improved by incorporating cohort-derived features, as measured by receiver operating characteristic (ROC) area under the curve (AUC) (Fig. 4B-D).

Next, we turned to evaluate a deep convolutional neural model that extracts information directly from the raw embryo images. These “deep learning” models were shown to surpass more traditional machine learning models in many domains, including IVF embryo implantation prediction [13,14,15,16]. Specifically, we used a pre-trained VGG16 network [53] and fine-tuned it using preprocessed images of transferred blastocysts. Here too, we trained one model without cohort features, and another with the network's confidence score along with cohort features (Methods). Similar to the previous models, inclusion of cohort features enhanced the model's capacity to accurately predict implantation outcome (Fig. 4E). Moreover, cohort information enhanced the capacity to accurately predict implantation for a model that combined the deep learning model score, morphology features and the oocyte age (Fig. 4F), as well as for a model that also included morphokinetic features (Fig. 4G). Finally, we validated that these results were consistent by performing 10-fold cross validation: 10 rounds of training and evaluation for each model, each time with an independent partitioning of cohorts to train and test sets (Fig. 4H).

These results, consistent across five models and multiple replicates, established that the siblings of the cohort encapsulate valuable information regarding the implantation potential of the transferred embryo.

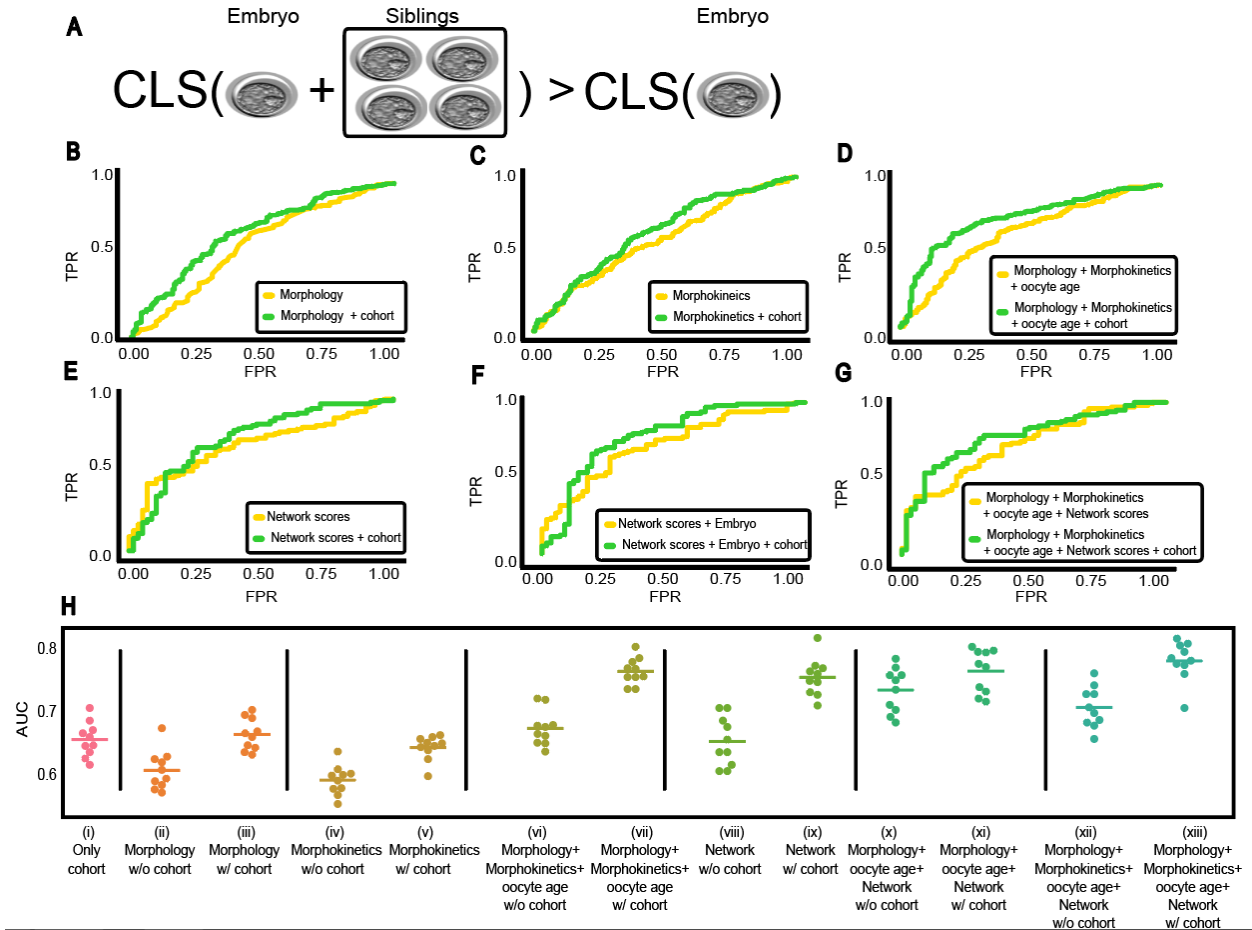


Figure 4: Implantation prediction performance (AUC) comparison of different model pairs with versus without sibling features establish that cohort properties contribute to implantation prediction. Statistical evaluation was computed with Wilcoxon signed rank test rejecting the null hypothesis that both models' predictions are drawn from the same distribution. **(A)** Illustration: do cohort siblings data contribute to implantation prediction? **(B)** Morphology (Gardner scores). N=1936 embryos. N positive embryos = 1141, N negative embryos = 795. AUC: 0.6 versus 0.68, respectively, p-value < 0.0001. **(C)** Morphokinetics. N=2089 embryos. N positive embryos = 1176, N negative embryos = 913. AUC: 0.591 and 0.641, respectively, p-value < 0.0001. **(D)** Morphokinetics, morphology and oocyte age: N=1936 embryos. N positive embryos = 1141, N negative embryos = 795. AUC: 0.662 versus 0.764, respectively, p-value < 0.0001. **(E-G)** Deep convolutional neural network without (E) and with (F) morphology and oocyte age, and with morphokinetics, morphology and oocyte age (G). N=772 embryos. N positive embryos = 482, N negative embryos = 290 (E) AUC: 0.698 versus 0.744, respectively, p-value < 0.001. (F) AUC: 0.72 versus 0.779, respectively, p-value < 0.001. (G) AUC: 0.727 versus 0.8, respectively, p-value < 0.001. **(H)** Replication analysis. Performance assessment for models trained without versus with cohort features in 10-folds cross validation. (i) Cohort features. AUC mean (standard deviation): 0.65 (0.02) (ii-iii) Morphology without (ii) or with (iii) cohort features. AUC mean (standard deviation)

was 0.59 (0.02) without cohort versus 0.65 (0.02) with cohorts, Wilcoxon signed rank test p-value < 0.01. (iv-v) Morphokinetics without (iv) or with (v) cohort features. AUC mean (standard deviation) was 0.57 (0.02) without cohort versus 0.62 (0.01) with cohort, Wilcoxon signed rank test p-value < 0.01. Surprisingly, the model trained with morphokinetic and cohort features performed slightly worse than the model trained with cohort features alone (i versus v) perhaps due to inclusion of morphological features less correlative with the outcome without feature selection. (vi-vii) Morphokinetics, morphology and oocyte age without (vi) or with (vii) cohort features. AUC mean (standard deviation) was 0.65 (0.02) without cohort versus 0.74 (0.02) with cohort, Wilcoxon signed rank test p-value < 0.01. (viii-ix) Deep convolutional neural network scores without (viii) or with (ix) cohort features. AUC mean (standard deviation) 0.63 (0.03) without cohort versus 0.73 (0.02) with cohort, Wilcoxon signed rank test p-value < 0.01. (x-xi) Deep convolutional neural network scores with morphology and oocyte age without (x) or with (xi) cohort features. AUC mean (standard deviation) was 0.71 (0.03) without cohort versus 0.75 (0.03) with cohort, Wilcoxon signed rank test p-value < 0.01. (xii-xiii) Deep convolutional neural network scores with morphokinetics, morphology and oocyte age without (xii) or with (xiii) cohort features. AUC mean (standard deviation) was 0.69 (0.03) without cohort versus 0.77 (0.02) with cohort, Wilcoxon signed rank test p-value < 0.01.

Identifying cohort properties driving the model's prediction

While cohort information was found to enhance implantation prediction, it was not clear which of the cohort features contributed to the measured boost in performance. Thus, we next aimed toward explaining the models' decisions. We focused our efforts on the two top performing models trained without or with the deep learning network's feature (i.e., confidence score), namely, a model trained on morphology, morphokinetics and oocyte age. This enabled us, beyond plain model interpretability, to assess what information was encapsulated in the network beyond the morphology, morphokinetics and oocyte age.

To pinpoint what features were the most important for the model's prediction we applied SHapley Additive exPlanation (SHAP), a game theory-based method for interpreting models' predictions that assigns each feature an importance value for a prediction [54]. For the model that was trained solely with cohort features we found that the fraction of siblings reaching blastulation and the cohort size were the two most important features for implantation prediction (Fig. 5A). The trophoctoderm score and expansion degree (morphology), oocyte age and seven morphokinetic features were identified as the ten top features in a model trained with morphokinetics, morphology and oocyte age (Fig. 5B). When comparing the ten top ranked features with a model trained with cohort features, we noticed that the fraction of sibling embryos reaching blastulation came up as the most important feature for prediction of implantation outcome, preceding morphology, morphokinetic and oocyte age (Fig. 5C). Overall,

five of the top ten important features were attributed to the cohort and the other five were attributed to the transferred blastocyst (Fig. 5C). The network confidence score was the most informative feature, by a margin, in a model trained with the network confidence score, morphology (Gardner scores), morphokinetics and the oocyte age (Fig. 5D). Trophectoderm score was ranked only third and the expansion degree was not ranked within the ten most important features indicating that the network encoded the annotated morphology (Gardner scores). Cohort features were ranked higher than the morphology and morphokinetic features in a model that was trained with the network prediction, morphology, morphokinetics and the oocyte age (Fig. 5E). Half of the top features (5/10) were cohort-related suggesting that the cohort encodes information that is not included within the embryo's image (Fig. 5E). While oocyte age was identified as an important feature in the two models that did not include cohort features (Fig. 5B and Fig. 5D), it was not one of the top ten features when cohort features were included (Fig. 5C and Fig. 5E). This suggests that the cohort encodes information more discriminative than the oocyte age in terms of implantation prediction. Altogether, these analyses show that cohort features are an important source of information for the prediction of implantation outcome.

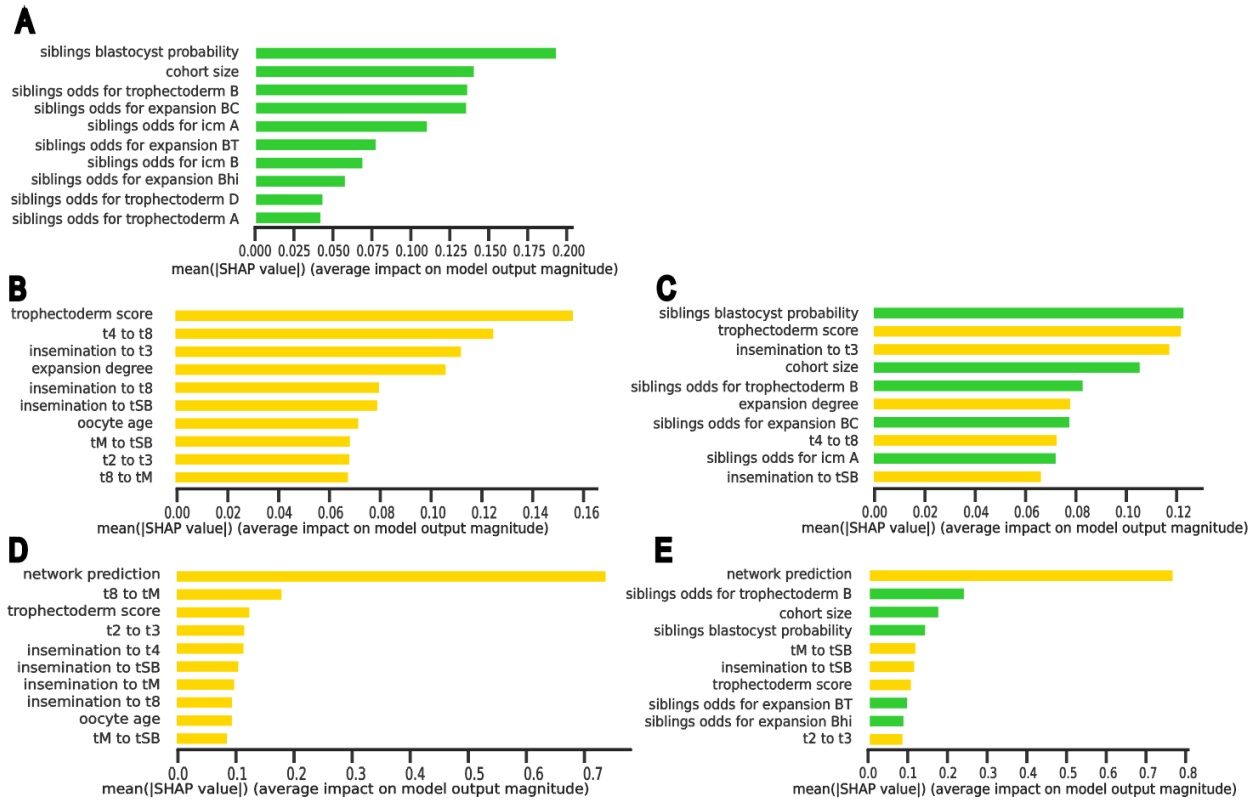


Figure 5: Model explainability analysis. Each panel shows features importance for the top ten features for a specific model using Shapely Additive Explanations (ShAP). Embryo features are marked in yellow and cohort features are marked with green. (A) Only cohort features. (B-C) Morphokinetics, morphology and oocyte age without (B) or with (C) cohort features. (D-E) Deep convolutional neural network predictions with morphokinetics, morphology, oocyte age and without (D) or with (E) cohort features.

Identifying cohort properties that corrected erroneous prediction

Finally, we evaluated the contribution of the cohort features to the classification of each of the transferred embryos. Fig. 6A shows the embryo classification scores by a model trained with morphology, morphokinetics and oocyte age without (x-axis) and with (y-axis) cohort features. Each data point corresponds to an embryo and the color code indicates positive (green) and negative (red) embryos. Embryos above the $y = x$ diagonal had higher classification scores when including the cohort features, reflecting a higher prediction for successful implantation. Inclusion of cohort features increased the classification scores of positive embryos (green data points above the $y = x$ diagonal) and decreased the classification scores of negative embryos (red data points below the $y = x$ diagonal), indicating that adding cohort features to the model improves model's discrimination for both positive and negative embryos (Fig. 6B). To better understand

how cohort features enhanced implantation prediction we zoomed in to the embryos that were “rescued” by the cohort features, i.e., correctly classified only by the classifier that had access to cohort information. These included 15 negative and 62 positive “rescued” embryos. We examined the two top ranked cohort features, fraction of sibling embryos reaching blastulation and cohort size. We compared these features in cohorts of “rescued” embryos to all cohorts with the label corresponding to the “erroneous” prediction (by the model that did not have access to the cohort). For example, the two aforementioned cohort features of a positive embryo that was predicted as “negative” by a model that did not have access to cohort information and “rescued” by the cohort features, were compared to the corresponding features of all negative cohorts. This analysis should highlight whether these cohort features were correlated with the “rescued” prediction, and thus provide insight on what cohort features are used to improve the model’s prediction. While not identifying an obvious pattern for rescued negative embryos (Fig. 6C-D), we revealed in rescued positive embryos an elevation (relative to negative embryos) in the fraction of sibling embryos reaching blastulation and in the cohort size (Fig. 6E-F). A similar pattern, although less reliable due to the lower number of rescued embryos (because of the smaller dataset), was observed for models that included the neural network’s predictions (Fig. S2). These results suggest that these two cohort features were used by the model to correct negative-to-positive predictions but not positive-to-negative predictions.

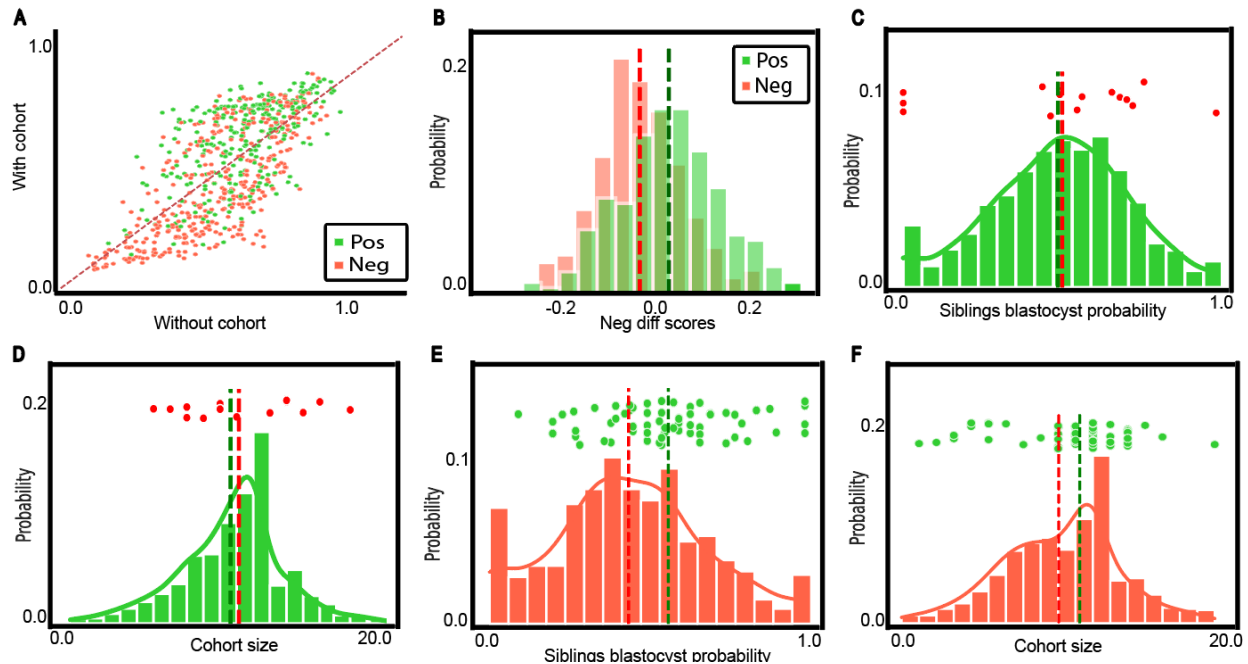


Figure 6: Analysis of cohort properties that “rescued” erroneous prediction. N transferred blastocysts = 2089 from which 1176 were positive and 913 were negative embryos. The results refer to a model trained with morphology, morphokinetics and oocyte age without and with cohort features. **(A)** Embryos matched classification scores by the two models, without (x-axis) and with (y-axis) cohort features. **(B)** Distribution of the difference in the embryos matched classification scores: with - without cohort features. Mean (standard deviation) difference for positive cohorts was 0.043 (0.1) (Wilcoxon signed rank test p-value < 0.0001) versus -0.02 (0.09) (Wilcoxon rank-sum test p-value < 0.01) for negative cohorts. **(C-F)** Distribution of fraction of blastocysts siblings (C,E) or cohort size (D,F) for positive (green, C-D) or negative (red, E-F) embryos. Each of the data points above the distribution indicate an embryo that was “rescued” with the cohort feature, i.e., classified erroneously by a model trained without and corrected with a model trained with cohort features. **(C-D)** Negative embryos that were erroneously classified as positive without cohort features and were correctly classified by a model that had access to cohort features. $N = 15$ rescued embryos. **(C)** Mean (standard deviation) fraction of sibling embryos within a cohort (not including the transferred embryo/s) reaching blastulation was 0.49 (0.22) for positive cohorts versus 0.5 (0.29) for negative rescued embryos, Wilcoxon signed rank test on the differences from the positive embryos’ mean was not statistically significant. **(D)** Mean (standard deviation) cohort size was 10.56 (3.19) for positive cohorts versus 11.3 (3.64) for negative rescued embryos, Wilcoxon signed rank test on the differences from the positive embryos’ mean was not statistically significant. **(E-F)** Positive embryos that were erroneously classified as negative without cohort features and were correctly classified by a model that had access to cohort features. $N = 62$ rescued embryos. Distribution of the fraction of embryos within a cohort (not including the transferred embryo/s) to develop to a blastocyst (E) or cohort size (i.e., number of sibling embryos in a cohort) (F) compared across negative embryos versus positive embryos that were “rescued” by the cohort features, i.e., correctly classified only by the classifier that had access to cohort information. **(E)** Mean (standard deviation) fraction of sibling embryos within a cohort (not including the transferred embryo/s) reaching blastulation was 0.41 (0.25) for negative cohorts versus 0.56 (0.21) for positive rescued embryos, Wilcoxon signed rank test on the differences from the negative embryos’ mean p-value < 0.0001. **(F)** Mean (standard deviation) cohort size was 9.9 (3.36) for negative

cohorts versus 10.74 (3.17) for positive rescued embryos, Wilcoxon signed rank test on the differences from the negative embryos' mean p-value < 0.01.

Discussion

In vitro fertilization (IVF) is a perfect system for studying the effect of genotypic and environmental variation on phenotype. This is due to the availability of high-content human embryo data that includes phenotypic information regarding multiple sibling embryos for each treatment cycle, who share a common genetic background and similar external conditions. From the machine learning perspective, IVF is a fitting example for an application where vast unlabeled data, specifically from non-transferred cohort siblings, can provide valuable information for a more accurate prediction of embryo phenotypic quality, i.e., implantation potential. These biological and machine learning concepts converge to a common theme where the uncertainty in the transferred embryo features, due to either inconsistency in annotations, features that were not explicitly measured or label ambiguity, can be reduced by information encapsulated in the correlated cohort embryos. We believe that this is achieved by noise reduction with the multiple correlated instances.

We established that embryos from the same cohort were more phenotypically similar than embryos from different cohorts (Fig. 2), demonstrated that siblings of successfully implanted (positive) embryos were of higher phenotypic quality in relation to siblings of negative embryos (Fig. 3), and demonstrated that cohort features contribute to machine learning based implantation prediction (Fig. 4). The latter was achieved by extracting a new set of features from unlabeled siblings within the cohort, incorporating them with different feature sets, and comparing the classifier's performance in implantation prediction without versus with cohort features. Even though each individual cohort feature had only a marginal effect (Fig. 3), the machine learning driven integration of all cohort features led to a consistent improvement in implantation prediction regardless of the embryo-focused model (Fig. 4). These results suggest a general concept where the transferred embryo's siblings encapsulate discriminative information that is complementary to the information encoded in the transferred embryo, and thus, cohort features are likely to contribute to any embryo-derived features. Since the siblings' data are routinely collected in the clinic, incorporating cohort features in AI-driven embryo implantation prediction can have direct translational implications in the clinic.

Previous studies correlated cohort-based properties to implantation outcome. For example, demonstrating improved outcomes for embryos selected from cohorts with more than five

embryos [25], or from day 3 cohorts where at least one sibling embryo achieved blastulation after extended culture [28,29]. Other studies incorporated specifically designed cohort-based features to machine learning models, specifically the cohort size [24,27] and number of developed embryos [38], or even incorporated multiple cohort-related features to show that the cohort alone contains discriminative information [26]. We performed a comprehensive analysis that systematically assessed the contribution of incorporation of cohort properties to existing models for prediction of implantation outcome. We did not decide on specific thresholds, rather we provided all the “raw” features to the AI machinery to automatically determine and weight which and what combination was most discriminative. This unbiased approach allowed us to reveal how discriminative each cohort property is in the context of a given model (Fig. 5) and which cohort properties were critical to “rescue” embryos that were incorrectly classified without cohort properties (Fig. 6). Importantly, most of the previous studies mentioned above assessed embryos that were transferred at day 3 from fertilization. At this early stage of embryo development, the uncertainty of the implantation potential of a single embryo is higher [30,31], and thus, additional correlated measurements in cohort features are expected to provide a more discriminative signal. This is especially relevant when sibling embryos are kept in extended blastocyst culture and their blastulation outcomes are known. Our results relate to blastocysts, i.e., day 5 embryos, when the uncertainty is lower. Still, we were able to establish that the cohort information contributes discriminative power beyond the transferred embryo features.

We found that cohort features “rescued” 4-fold more positive ($N = 62$) versus negative ($N = 15$) embryos (Fig. 6). One explanation for this asymmetry could be due to the ambiguity in the negative labels. While positive embryos are inherently of high implantation quality, negative embryos fail implantation because of either low quality embryos and/or poor endometrial or uterine factors. This creates uncertainty in the ground truth labels of non-implanted embryos, or “label ambiguity” [20]. Thus, it could be easier to “rescue” a positive embryo that was mistakenly classified as “negative” with a quality cohort, in comparison to a negative embryo that could be of high implantation potential (also encoded in its cohort features).

In this study, digital embryo images were manually annotated for morphology (based on the Gardner embryo scoring system) and key morphokinetic events using the time-lapse information. Manual annotation is throughput-limiting. Automated tools for morphological evaluation [22,23] and detection of morphokinetic events [20,55] are quickly reaching human-level performance

[56] and have the advantage of avoiding intra- and inter-annotator bias potentially replacing the need for manual annotations in the near future.

In the context of machine learning, this is an example of semi-supervised learning, where a small fraction of labeled (transferred embryos) and vast unlabeled observations (non-transferred cohort siblings) are used together to improve the learned model's performance toward the task of predicting implantation potential. While the underlying assumption in semi-supervised learning is that the observations are unrelated, here we characterize another type of semi-supervised learning, where the unlabeled observations are associated with the labeled observations, and suggest a way to use information from these unlabeled observations for improving the prediction power of the supervised models.

Methods

Experiments

Data collection and ethics

The data included for this analysis were retrospectively collected from IVF cycles conducted at IVI Valencia fertility center, Spain, between March 2010 and December 2018. Historical images of blastocyst-stage embryos and metadata were provided by AiVF. All procedures and protocols were approved by an Institutional Review Board for secondary research use (IRB reference number HMO-006-20). All IVF cycles were with either the patient's own oocytes (n=1134) or with donor oocytes (n=955), ages 18-51 years old (Fig. S3). In cases of oocyte donation, donor age was considered as the “oocyte age”.

Fertilization was determined by the presence of two pronuclei (2PN) 16-18 hours after insemination. All zygotes were placed inside the EmbryoScope™ time-lapse incubator system (Vitrolife, Denmark) and incubated using sequential media protocol until blastocyst-stage. Image acquisition using the EmbryoScope™ imaging software occurred every 15-20 minutes. For every embryo, seven-layers of z-stack images 15µm apart were acquired at each time point, where time 0 was defined as the fertilization time. A total of 16,194 oocytes from 1,605 IVF cycles were recorded. All cycles included at least one fresh single embryo transfer at blastocyst stage with a known implantation outcome (implanted/not implanted). Out of 2089 embryos with known implantation outcomes, 1176 were successfully implanted (*positive embryos*) and 913 failed to implant (*negative embryos*).

Annotation of embryo clinical quality

Following in vitro incubation, the number of embryos from the cohort transferred was either one or two. In cycles where two embryos were transferred, we only included cases where both embryos either successfully implanted or both failed implantation. Implantation following embryo transfer was determined by ultrasound scanning for gestational sac after ~seven weeks of pregnancy. Positive embryos (and their corresponding positive cohorts) were defined when the number of gestational sacs matched the number of transferred embryos. Negative embryos (and

their corresponding negative cohorts) were defined when no gestational sac following embryo transfer was observed.

Analysis

Embryo morphological feature representation

Embryos were morphologically annotated at the blastocyst stage (day 5 of embryonic development) according to the Gardner scoring scheme by onsite trained embryologists ~120 hours post insemination [8,50]. The time of blastulation was determined by time-lapse monitoring of embryo development. Specifically, every embryo was assigned a three-part alphanumeric quality score based on its expansion status (“blastocyst expansion”, ranked 1-6), morphology of the inner cell mass (“ICM”, ranked A-D), and morphology of the trophectoderm (“TE”, ranked A-D) (Fig. 2H). Embryos with missing Gardner annotations were excluded from the analysis. Out of the 2089 embryos, 1936 had all Gardner annotations, where 1141 were annotated as positive and 795 as negative. Embryos’ morphological variables were computationally represented via one-hot encoding, i.e., a feature vector of size $5 + 4 + 4 = 13$ representing the Gardner scores.

Embryo morphokinetic feature representation

Time-lapse images of embryo development were viewed by onsite trained embryologist and seven key morphokinetic events were manually annotated in accordance with published consensus criteria [57]: Time of fertilization (t_0), cell division to the 2,3,4, and 8-cell stage (t_2 , t_3 , t_4 , t_8), compaction of the morula (t_M), and time of blastulation (t_{SB}). Missing annotations due to limitations of the dataset were completed according to the following rules, as determined by domain experts: $t_2 =$ time of pronuclei disappearance (t_{PNf} - time when both pronuclei disappear, independently annotated, see below) + 2 hours; $t_3 = t_4 - 1$ hour; $t_4 = t_3 + 1$ hour; $t_8 = t_7 + 3$ hours; $t_M = t_{SB} - 6$ hours. Missing morphokinetic event annotations were at levels of 5%-10%, except for t_M and t_{SB} with 20% missing annotations. After manually completing the missing annotations, embryos with a further single missing value were determined by first finding the five most similar embryos based on the remaining available annotated morphokinetic

features and then using their mean value of the missing features. Overall, the morphokinetic feature vector was of size 11 and included the (five) time intervals between every consecutive developmental stage and the (six) overall time points for each developmental stage from the time of fertilization. Importantly, our dataset also included manual annotations beyond the morphokinetic events listed above (e.g., tPNf, see above). These events were not included in the morphokinetic feature representation used in this analysis because many embryos in the dataset did not have these morphokinetic events annotated.

Similarity between timing of morphokinetic events

We compared the similarity between the timings of morphokinetic events and between the timings of consecutive morphokinetic events between embryo pairs within the same cohort (siblings) and in different cohorts (non siblings). For each morphokinetic event timing and consecutive morphokinetic events timing, we compared the similarity of all sibling and non-sibling pairs. A similarity measure that encodes the full morphokinetic profile was calculated as the Euclidean distance between normalized vectors that included the timing of all the morphokinetic events and the timing of all the consecutive morphokinetic events.

To compare morphokinetic similarities between sibling and non sibling embryo pairs with similar morphological properties, we evaluated all embryo triplets that share the same Gardner scores, where two embryos were siblings and the third embryo was from a different cohort. In each triplet we evaluated the similarity in morphokinetics between the siblings versus the non siblings pairs.

Cohort morphology-based feature representation

A cohort contains multiple embryos from the same couples in the same IVF treatment cycle. Seventeen (17) cohort features were extracted from all cohort siblings, excluding the transferred embryo. These included the cohort size (the only feature that included the transferred embryo), the fraction of sibling embryos reaching blastulation (Fig. 3B), the fraction of sibling embryos that hatched, a 13-dimensional vector encoding the fraction of siblings for each Gardner score (5+4+4 features).

Automated deep learning based embryo implantation prediction

We had access to historical images of 772 transferred blastocyst-stage embryos from 638 cohorts and associated known implantation outcomes. Of these, 482 embryos successfully implanted (positive) and 290 failed to implant (negative).

Images Preprocessing

Our analysis focused on the blastocyst's last time-frame prior to hatching and the center z-stack image. We preprocessed the raw frames in order to segment the embryo region from the surrounding well and background. This allowed us to train more complex models on a large training set in reasonable time while keeping the inherent spatial resolution features of the embryo. To localize and segment the embryo we developed the following pipeline. First, a mask-RCNN [58] was trained to identify a bounding box around each embryo using 800 manually annotated images. Second, a hough-transform [59] was applied to center the embryo in the image by detecting a circular object within the bounding box. Third, a U-NET model was trained based on 500 manually validated outputs from the previous step to provide the embryo segmentation mask [60]. Our U-NET architecture consisted of 4 convolutional layers for the encoder (downsampling) / decoder (upsampling) with 32, 64, 128 and 256 filters correspondingly. Each layer included batch normalization and relu activation function with maxpooling for the encoder and upsampling for the decoder. Finally, the image was further resized to 64x64. The trained pipeline architecture is presented in Fig. S4.

Image classification model

A pretrained VGG16 [53] architecture was used as a backbone followed by a flattening layer, a fully connected 16 node dense layer and a dense single node layer with a sigmoid activation. The full model was retrained (no freezing of weights) using binary cross entropy loss and the Adam optimizer with a learning rate of 0.00001 to output a confidence score in the range of 0-1 for predicting implantation probability.

The model was trained for 100 epochs with a batch size of 32. For each batch, image-augmentation (brightness, flipping, rotation and Gaussian noise) was performed prior to fitting the model to reduce overfit. We trained the model with 10-fold cross validation allowing all the available data points to run at inference and output a confidence score.

Evaluation of machine learning models without versus with cohort features

Models were trained with morphology features and/or morphokinetic features and/or the oocyte age. Each feature representation was used to train and then evaluate the performance of two models: without including cohort features (i.e., using only the transferred embryo information) versus including cohort features. The deep learning model was used as a feature extractor, where the model confidence score was one feature, and this features were used together with other features (e.g., morphokinetics, oocyte age) to train and then evaluate the performance of models that had vs. those did not have access to cohort features. Models based on the same features without or with cohort features were compared to evaluate the contribution of cohort features on the overall prediction. The data was partitioned to train (80%) and test (20%) sets maintaining a similar ratio of positive (i.e., successful implantation) and negative (failed implantation) cohorts in both sets. An XGBoost binary classification model was trained for each feature set [61]. When included, the confidence score of the deep learning model was considered as a single feature. Receiver Operating Characteristic (ROC) curve (i.e., true positive rate (TPR, sensitivity) versus false positive rate (FPR, 1-specificity)) and its Area Under the Curve (AUC) were used to visually and quantitatively compare performance of model pairs without versus with cohort features. Hyper-parameter tuning was performed independently for each XGBoost classifier using sklearn's GridSearchCV [62]. The parameters that were optimized were: (1) Number of gradient boosted trees, (2) Minimum loss reduction required to make a further partition on a leaf node of the tree, (3) Maximum tree depth for base learners, (4) Subsample ratio of the training instance, (5) Subsample ratio of columns when constructing each tree, and (6) Minimum sum of instance weight (hessian) needed in a child.

Due to the lower volume of annotated raw image data, we applied feature selection, independently for each model, with the extremely Randomized Trees Classifier (ExtraTrees) classifier (an updated version of random forest) [63] to reduce the number of features to ten (unless the number of features was already lower).

Feature importance analysis

For interpretability purposes, we applied the feature importance method with Shapley Additive ExPlanations (SHAP) [54]. SHAP computes the contribution of each feature to each individual

prediction. Embryo morphological categorical parameters were transformed into ordinal variables for the SHAP analysis.

References

1. Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., & Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, 2(3), e41. <https://doi.org/10.1371/journal.pgen.0020041>
2. Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P. I., Godneva, A., Kalka, I. N., Bar, N., Shilo, S., Lador, D., Vila, A. V., Zmora, N., Pevsner-Fischer, M., Israeli, D., Kosower, N., Malka, G., Wolf, B. C., ... Segal, E. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature*, 555(7695), 210–215. <https://doi.org/10.1038/nature25973>
3. Jirtle, R. L., & Skinner, M. K. (2007). Environmental epigenomics and disease susceptibility. *Nature Reviews. Genetics*, 8(4), 253–262. <https://doi.org/10.1038/nrg2045>
4. Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., Heine-Suñer, D., Cigudosa, J. C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T. D., Wu, Y.-Z., ... Esteller, M. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30), 10604–10609. <https://doi.org/10.1073/pnas.0500398102>
5. Sang, Q., Zhou, Z., Mu, J., & Wang, L. (2021). Genetic factors as potential molecular markers of human oocyte and embryo quality. *Journal of Assisted Reproduction and Genetics*, 38(5), 993–1002. <https://doi.org/10.1007/s10815-021-02196-z>
6. Gardner, D. K., & Kelley, R. L. (2017). Impact of the IVF laboratory environment on human preimplantation embryo phenotype. *Journal of Developmental Origins of Health and Disease*, 8(4), 418–435. <https://doi.org/10.1017/S2040174417000368>
7. Meseguer, M., Herrero, J., Tejera, A., Hilligsoe, K. M., Ramsing, N. B., & Remohi, J. (2011). The use of morphokinetics as a predictor of embryo implantation. *Human Reproduction*, 26(10), 2658–2671. <https://doi.org/10.1093/humrep/der256>
8. Gardner, D. K., & Schoolcraft, W. B. (1999). Culture and transfer of human blastocysts. *Current Opinion in Obstetrics & Gynecology*, 11(3), 307–311. <https://doi.org/10.1097/00001703-199906000-00013>

9. Le Nhat Quang, le Thi Bich Tram, Nguyen Huyen Minh Thuy, & Toan Pham. (2018). Comparison of clinical outcome of frozen embryo transfer after embryo selection based on morphokinetic versus morphologic criteria for freezing. *Biomedical Research and Therapy*, 5(12), 2910–2917.
10. Coticchio, G., Mignini Renzini, M., Novara, P. v, Lain, M., de Ponti, E., Turchi, D., Fadini, R., & Dal Canto, M. (2018). Focused time-lapse analysis reveals novel aspects of human fertilization and suggests new parameters of embryo viability. *Human Reproduction (Oxford, England)*, 33(1), 23–31. <https://doi.org/10.1093/humrep/dex344>
11. Filho, E. S., Noble, J. A., & Wells, D. (2010). A review on automatic analysis of human embryo microscope images. *The Open Biomedical Engineering Journal*, 4, 170–177. <https://doi.org/10.2174/1874120701004010170>
12. Adolfsson, E., Porath, S., & Andershed, A. N. (2018). External validation of a time-lapse model; a retrospective study comparing embryo evaluation using a morphokinetic model to standard morphology with live birth as endpoint. *JBRA Assisted Reproduction*, 22(3), 205–214. <https://doi.org/10.5935/1518-0557.20180041>
13. Raef, B., & Ferdousi, R. (2019). A Review of Machine Learning Approaches in Assisted Reproductive Technologies. *Acta Informatica Medica : AIM : Journal of the Society for Medical Informatics of Bosnia & Herzegovina : Casopis Drustva Za Medicinsku Informatiku BiH*, 27(3), 205–211. <https://doi.org/10.5455/aim.2019.27.205-211>
14. Simopoulou, M., Sfakianoudis, K., Maziotis, E., Antoniou, N., Rapani, A., Anifandis, G., Bakas, P., Bolaris, S., Pantou, A., Pantos, K., & Koutsilieris, M. (2018). Are computational applications the “crystal ball” in the IVF laboratory? The evolution from mathematics to artificial intelligence. *Journal of Assisted Reproduction and Genetics*, 35(9), 1545–1557. <https://doi.org/10.1007/s10815-018-1266-6>
15. Malinowski, P., Milewski, R., Ziniewicz, P., Milewska, A. J., Czerniecki, J., & Wołczyński, S. (2014). The Use of Data Mining Methods to Predict the Result of Infertility Treatment Using the IVF ET Method. *Studies in Logic, Grammar and Rhetoric*, 39(1), 67–74. <https://doi.org/10.2478/slgr-2014-0044>
16. Tran, D., Cooke, S., Illingworth, P. J., & Gardner, D. K. (2019). Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst

- transfer. *Human Reproduction (Oxford, England)*, 34(6), 1011–1018.
<https://doi.org/10.1093/humrep/dez064>
17. Uyar, A., Bener, A., & Ciray, H. N. (2015). Predictive Modeling of Implantation Outcome in an In Vitro Fertilization Setting: An Application of Machine Learning Methods. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, 35(6), 714–725. <https://doi.org/10.1177/0272989X14535984>
 18. David H. Silver, Martin Feder, Yael Gold-Zamir, Avital L. Polsky, Shahar Rosentraub, Efrat Shachorn, Adi Weinberger, Pavlo Mazur, Valery D. Zukin, & Alex M. Bronstei. (2020). Data-Driven Prediction of Embryo Implantation Probability Using IVF Time-lapse Imaging. *ArXiv*, 2006(01035).
 19. Khosravi, P., Kazemi, E., Zhan, Q., Malmsten, J. E., Toschi, M., Zisimopoulos, P., Sigaras, A., Lavery, S., Cooper, L. A. D., Hickman, C., Meseguer, M., Rosenwaks, Z., Elemento, O., Zaninovic, N., & Hajirasouliha, I. (2019). Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ Digital Medicine*, 2, 21. <https://doi.org/10.1038/s41746-019-0096-y>
 20. Itay Erlich, Assaf Ben-Meir, Iris Har-Vardi, James A. Grifo, & Assaf Zaritsky. (2021). Solving the “right” problems for effective machine learning driven in vitro fertilization. *MedRxiv*.
 21. Erlich, I., Ben-Meir, A., Har-Vardi, I., Grifo, J., Wang, F., Mccaffrey, C., McCulloh, D., Or, Y., & Wolf, L. (2022). Pseudo contrastive labeling for predicting IVF embryo developmental potential. *Scientific Reports*, 12(1), 2488.
<https://doi.org/10.1038/s41598-022-06336-y>
 22. Pegah Khosravi, & Ehsan Kazemi. (2018). Robust Automated Assessment of Human Blastocyst Quality using Deep Learning. *BioRxiv*.
 23. Kragh, M. F., Rimestad, J., Berntsen, J., & Karstoft, H. (2019). Automatic grading of human blastocysts from time-lapse imaging. *Computers in Biology and Medicine*, 115, 103494. <https://doi.org/10.1016/j.combiomed.2019.103494>
 24. Durairaj M, & Thamilselvan P. (2013). Applications of Artificial Neural Network for IVF Data Analysis and Prediction. *Journal of Engineering, Computers & Applied Sciences* .
 25. Devreker, F., Pogonici, E., de Maertelaer, V., Revelard, P., van den Bergh, M., & Englert, Y. (1999). Selection of good embryos for transfer depends on embryo cohort size:

- implications for the “mild ovarian stimulation” debate. *Human Reproduction (Oxford, England)*, 14(12), 3002–3008. <https://doi.org/10.1093/humrep/14.12.3002>
26. Jun, S. H., Choi, B., Shahine, L., Westphal, L. M., Behr, B., Reijo Pera, R. A., Wong, W. H., & Yao, M. W. M. (2008). Defining human embryo phenotypes by cohort-specific prognostic factors. *PloS One*, 3(7), e2562. <https://doi.org/10.1371/journal.pone.0002562>
27. Blank, C., Wildeboer, R. R., DeCruo, I., Tilleman, K., Weyers, B., de Sutter, P., Mischi, M., & Schoot, B. C. (2019). Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. *Fertility and Sterility*, 111(2), 318–326. <https://doi.org/10.1016/j.fertnstert.2018.10.030>
28. Fisch, J. D., Milki, A. A., & Behr, B. (1999). Sibling embryo blastocyst development correlates with the in vitro fertilization day 3 embryo transfer pregnancy rate in patients under age 40. *Fertility and Sterility*, 71(4), 750–752. [https://doi.org/10.1016/s0015-0282\(98\)00533-0](https://doi.org/10.1016/s0015-0282(98)00533-0)
29. Mackenna, A., Crosby, J., & Zegers-Hochschild, F. (2013). Sibling embryo blastocyst development as a prognostic factor for the outcome of day-3 embryo transfer. *Reproductive Biomedicine Online*, 26(5), 486–490. <https://doi.org/10.1016/j.rbmo.2013.02.003>
30. Alper, M. M., Brinsden, P., Fischer, R., & Wikland, M. (2001). To blastocyst or not to blastocyst? That is the question. *Human Reproduction (Oxford, England)*, 16(4), 617–619. <https://doi.org/10.1093/humrep/16.4.617>
31. Levitas, E., Lunenfeld, E., Har-Vardi, I., Albotiano, S., Sonin, Y., Hackmon-Ram, R., & Potashnik, G. (2004). Blastocyst-stage embryo transfer in patients who failed to conceive in three or more day 2-3 embryo transfer cycles: a prospective, randomized study. *Fertility and Sterility*, 81(3), 567–571. <https://doi.org/10.1016/j.fertnstert.2003.08.031>
32. Wang, S. S., & Sun, H. X. (2014). Blastocyst transfer ameliorates live birth rate compared with cleavage-stage embryos transfer in fresh in vitro fertilization or intracytoplasmic sperm injection cycles: reviews and meta-analysis. *Yonsei medical journal*, 55(3), 815-825.
33. Hardarson, T., Van Landuyt, L., & Jones, G. (2012). The blastocyst. *Human reproduction*, 27(suppl_1), i72-i91.

34. Curchoe, C. L., & Bormann, C. L. (2019). Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. *Journal of Assisted Reproduction and Genetics*, 36(4), 591-600.
35. Qazi, G. (2011). Obstetric and perinatal outcome of multiple pregnancy. *J Coll Physicians Surg Pak*, 21(3), 142-5.
36. Gleicher, N., Kushnir, V. A., & Barad, D. H. (2016). Risks of spontaneously and IVF-conceived singleton and twin pregnancies differ, requiring reassessment of statistical premises favoring elective single embryo transfer (eSET). *Reproductive Biology and Endocrinology*, 14(1), 1-7.
37. Hassan, M. R., Al-Insaif, S., Hossain, M. I., & Kamruzzaman, J. (2020). A machine learning approach for prediction of pregnancy outcome following IVF treatment. *Neural computing and applications*, 32(7), 2283-2297.
38. Hafiz, P., Nematollahi, M., Boostani, R., & Namavar Jahromi, B. (2017). Predicting Implantation Outcome of In Vitro Fertilization and Intracytoplasmic Sperm Injection Using Data Mining Techniques. *International Journal of Fertility & Sterility*, 11(3), 184–190. <https://doi.org/10.22074/ijfs.2017.4882>
39. Guh, R. S., Wu, T. C. J., & Weng, S. P. (2011). Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes. *Expert Systems with Applications*, 38(4), 4437-4449.
40. Rubio, I., Galán, A., Larreategui, Z., Ayerdi, F., Bellver, J., Herrero, J., & Meseguer, M. (2014). Clinical validation of embryo culture and selection by morphokinetic analysis: a randomized, controlled trial of the EmbryoScope. *Fertility and sterility*, 102(5), 1287-1294.
41. Shapiro, B. S., Daneshmand, S. T., Garner, F. C., Aguirre, M., & Thomas, S. (2008). Large blastocyst diameter, early blastulation, and low preovulatory serum progesterone are dominant predictors of clinical pregnancy in fresh autologous cycles. *Fertility and sterility*, 90(2), 302-309.
42. Ahlström, A., Westin, C., Reismer, E., Wikland, M., & Hardarson, T. (2011). Trophoctoderm morphology: an important parameter for predicting live birth after single blastocyst transfer. *Human reproduction*, 26(12), 3289-3296.
43. Wu, L., Han, W., Zhang, X., Wang, J., Liu, W., Xiong, S., & Huang, G. (2016). A retrospective analysis of morphokinetic parameters according to the implantation outcome of IVF treatment. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 197, 186-190.
44. Rocha, C., Nogueira, M. G., Zaninovic, N., & Hickman, C. (2018). Is AI assessment of morphokinetic data and digital image analysis from time-lapse culture predictive of implantation potential of human embryos?. *Fertility and Sterility*, 110(4), e373.

45. Barash, O., Ivani, K., Weckstein, L., & Hinckley, M. (2018). High accuracy machine learning predictive model for embryo selection in IVF PGT cycles with single embryo transfers. *Fertility and Sterility*, 110(4), e372.
46. Bormann, C. L., Kanakasabapathy, M. K., Thirumalaraju, P., Gupta, R., Pooniwala, R., Kandula, H., ... & Shafiee, H. (2020). Performance of a deep learning based neural network in the selection of human blastocysts for implantation. *Elife*, 9, e55301.
47. Bormann, C. L., Thirumalaraju, P., Kanakasabapathy, M. K., Kandula, H., Souter, I., Dimitriadis, I., ... & Shafiee, H. (2020). Consistency and objectivity of automated embryo assessments using deep neural networks. *Fertility and sterility*, 113(4), 781-787.
48. Chamayou, S., Patrizio, P., Storaci, G., Tomaselli, V., Alecci, C., Ragolia, C., Crescenzo, C., & Guglielmino, A. (2013). The use of morphokinetic parameters to select all embryos with full capacity to implant. *Journal of Assisted Reproduction and Genetics*, 30(5), 703–710. <https://doi.org/10.1007/s10815-013-9992-2>
49. Gardner, D. K., Lane, M., Stevens, J., Schlenker, T., & Schoolcraft, W. B. (2000). Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertility and Sterility*, 73(6), 1155–1158. [https://doi.org/10.1016/s0015-0282\(00\)00518-5](https://doi.org/10.1016/s0015-0282(00)00518-5)
50. Taketo Inoue, Yukiko Yonezawa, Junji Kishi, Nobuyuki Emi, & Yoshiyuki Ono. (2014). Improvement of Live Birth Rate Following Frozen-Thawed Blastocyst Transfer by Combination of Prednisolone Administration and Stimulation of Endometrium Embryo Transfer. *Open Journal of Obstetrics and Gynecology*.
51. Yan, J., Wu, K., Tang, R., Ding, L., & Chen, Z.-J. (2012). Effect of maternal age on the outcomes of in vitro fertilization and embryo transfer (IVF-ET). *Science China. Life Sciences*, 55(8), 694–698. <https://doi.org/10.1007/s11427-012-4357-0>
52. Padilla, S. L., & Garcia, J. E. (1989). Effect of maternal age and number of in vitro fertilization procedures on pregnancy outcome. *Fertility and Sterility*, 52(2), 270–273. [https://doi.org/10.1016/s0015-0282\(16\)60854-3](https://doi.org/10.1016/s0015-0282(16)60854-3)
53. Karen Simonyan, & Andrew Zisserman. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556*.
54. Lundberg, S. M., & Su-In Lee. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
55. Nir Zabari, Zeev Shoham, Yoel Shofaro, Dganit Richter, Iris Har-Vardi, Assaf Ben-Meir, Naama Srebnik, Amnon Buxboim, Yuval Or, & Yoav Kan-Tor. (2022). Delineating the

heterogeneity of preimplantation development via unsupervised clustering of embryo candidates for transfer using automated, accurate and standardized morphokinetic annotation. *MedRxiv*.

56. Nathan H, Julian McAuley, Julian A Gingold, Nina Desai, & Zachary C Lipton. (2018). Predicting Embryo Morphokinetics in Videos with Late Fusion Nets & Dynamic Decoders. *ICLR 2018 Workshop Submission*.
57. Ciray, H. N., Campbell, A., Agerholm, I. E., Aguilar, J., Chamayou, S., Esbert, M., Sayed, S., & Time-Lapse User Group. (2014). Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. *Human Reproduction (Oxford, England)*, 29(12), 2650–2660.
<https://doi.org/10.1093/humrep/deu278>
58. Kaiming He, Georgia Gkioxari, Piotr Dollár, & Ross Girshick. (2018). Mask R-CNN. *ArXiv:1703.06870*.
59. Arthur Coste. (2012). Image Processing : Hough Transform. *Computer Vision and Image Processing Course Work*.
60. Olaf Ronneberger, Philipp Fischer, & Thomas Brox. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv:1505.04597*.
61. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
62. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR*.
63. Geurts, P., Damien Ernst, & Louis Wehenkel. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 2–42.
64. Nigam, K., McCallum, A., & Mitchell, T. M. (2006). Semi-Supervised Text Classification Using EM.
65. Ranzato, M. A., & Szummer, M. (2008). Semi-supervised learning of compact document representations with deep networks. In *Proceedings of the 25th international conference on Machine learning* (pp. 792-799).

Supplementary Tables

	# embryos	# implanted embryos	# non-implanted embryos	# cohorts	# positive cohorts	# negative cohorts	# siblings
Morphokinetics annotated	2089	1176	913	1605	928	677	14105
Morphology annotated	1936	1141	795	1513	908	605	13493
Image data	772	482	290	638	404	234	5990

Table S1: Number of transferred blastocysts, number of implanted blastocysts (positive embryos) and non-implanted blastocysts (negative embryos), number of cohorts, number of positive and negative cohorts and number of cohort's siblings for each dataset.

Supplementary Figures

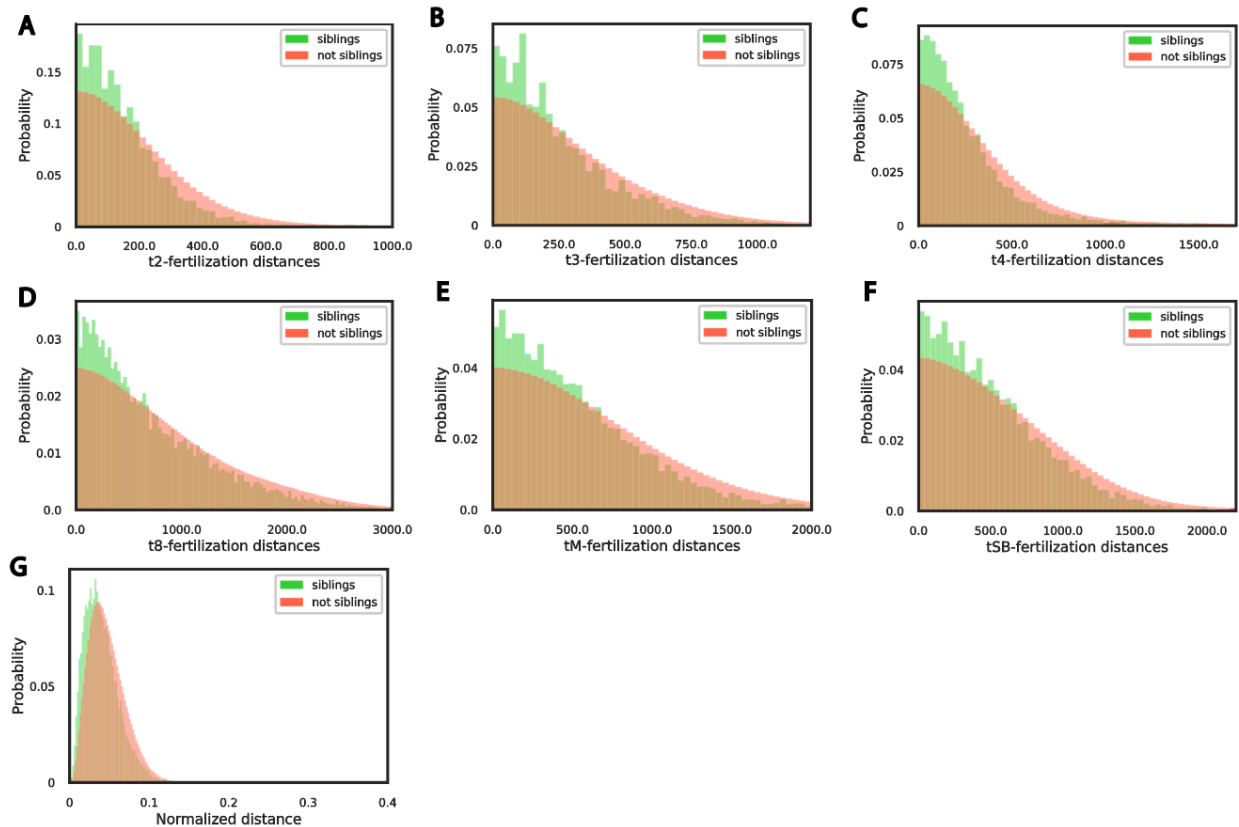


Figure S1. Sibling embryos from the same cohort are more similar than non-siblings. (A-F) Distribution of the difference in time intervals, in minutes (A-F) or normalized distance (G), between fertilization and each morphokinetic event compared across siblings versus not siblings embryo pairs. morphokinetic features: cell division to the 2, 3, 4 and 8-cell stage (t2, t3, t4, t8), the compaction of the morula - a day-3 development stage (tM) and the start of blastulation (tSB) - a day-5 development stage. N embryos = 16194. N cohorts = 1605. N positive cohorts = 928, N negative cohorts = 677. (A) Mean (standard deviation) of distances between fertilization-t2 intervals was 158.56 (150.59) for sibling embryos versus 217.62 (218.51) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (B) Mean (standard deviation) of distances between fertilization-t3 intervals was 258.21 (238.36) for sibling embryos versus 327.68 (290.56) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (C) Mean (standard deviation) of distances between fertilization-t4 intervals was 261.45 (268.56) for sibling embryos versus 333.01 (314.93) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (D) Mean (standard deviation) of distances between fertilization-t8 intervals was 700.46 (595.04) for sibling embryos versus 836.65 (666.99) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (E) Mean (standard deviation) of distances between fertilization-tM intervals was 537.38 (443.54) for sibling embryos versus 655.45 (515.67) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (F) Mean (standard deviation) of distances between fertilization-tSB intervals was 499.29 (397.18) for sibling embryos versus 596.34 (462.68) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001. (G) Mean (standard deviation) of normalized distances between all morphokinetic features time intervals from fertilization was 0.04 (0.02) for sibling embryos versus 0.05 (0.02) for not-sibling embryos, Mann-Whitney-U signed rank test p-value < 0.0001.

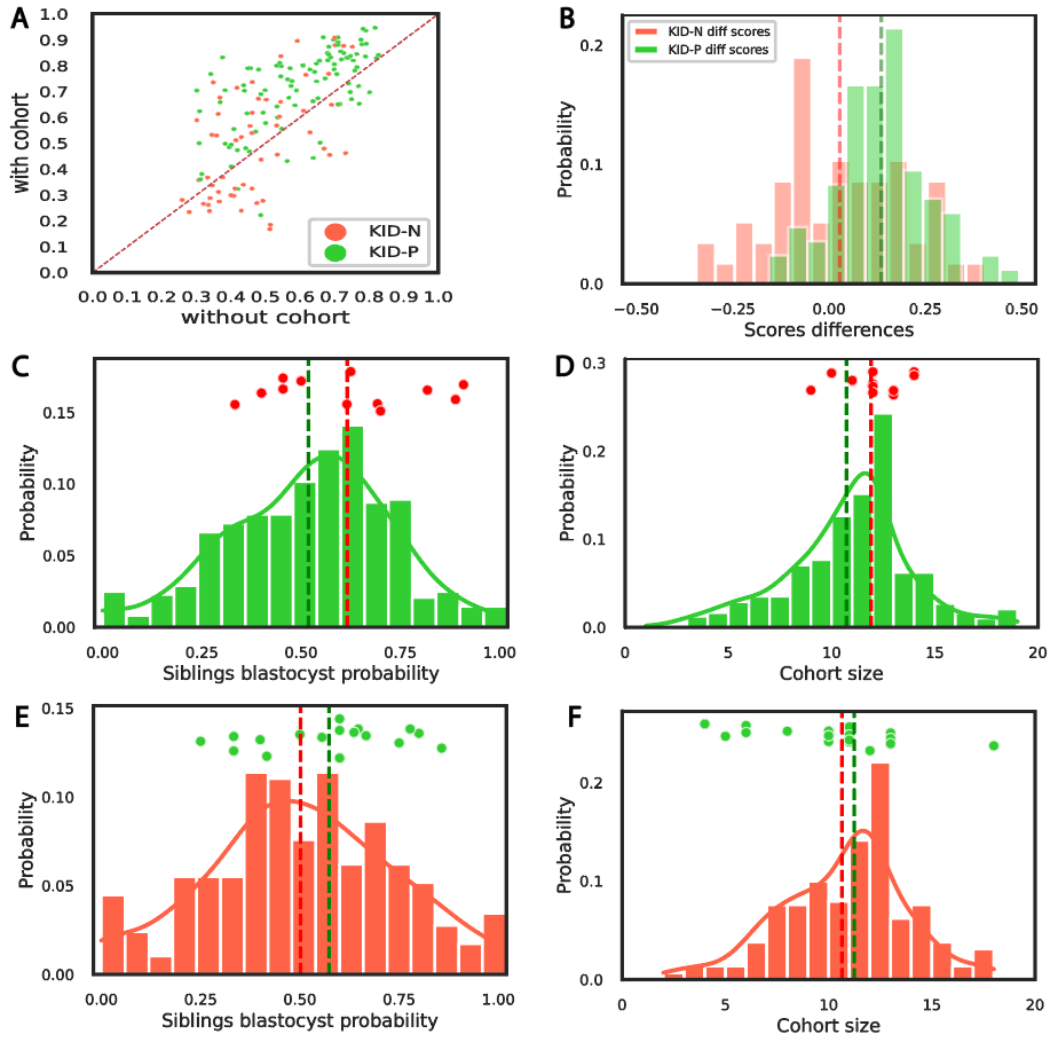


Figure S2: Analysis of cohort properties that “rescued” erroneous prediction. N transferred blastocysts = 772 from which 482 were positive and 290 were negative embryos. The results refer to a model trained with deep convolutional neural network, morphology, morphokinetics and oocyte age without and with cohort features. **(A)** Embryos matched classification scores by the two models: without (x-axis) and with (y-axis) cohort features. **(B)** Distribution of the difference in the embryos matched classification scores: with - without cohort features. Mean (standard deviation) difference for positive cohorts was 0.13 (0.11) (Wilcoxon rank-sum test p-value < 0.0001) versus 0.01 (0.17) (Wilcoxon rank-sum test was not statistically significant) for negative cohorts. **(C-F)** Distribution of fraction of blastocysts siblings (C,E) or cohort size (D,F) for positive (green, C-D) or negative (red, E-F) embryos. Each of the data points above the distribution indicate an embryo that was “rescued” with the cohort feature, i.e., classified erroneously by a model trained without and corrected with a model trained with cohort features. **(C-D)** Negative embryos that were erroneously classified as positive without cohort features and were correctly classified by a model that had access to cohort features. N = 12 rescued embryos. **(C)** Mean (standard deviation) fraction of sibling embryos within a cohort (not including the transferred embryo/s) reaching blastulation was 0.51 (0.2) for positive cohorts versus 0.61 (0.19) for negative rescued embryos, Wilcoxon signed rank test on the differences from positive embryos mean found no statistical significance. **(D)** Mean (standard

deviation) cohort size was 10.56 (3.19) for positive cohorts versus 11.91 (3.64) for negative rescued embryos, Wilcoxon signed rank test on the differences from positive embryos mean found no statistical significance. **(E-F)** Positive embryos that were erroneously classified as negative without cohort features and were correctly classified by a model that had access to cohort features. N = 18 rescued embryos. Distribution of the fraction of embryos within a cohort (not including the transferred embryo/s) to develop to a blastocyst (E) or cohort size (i.e., number of sibling embryos in a cohort) (F) compared across negative embryos versus positive embryos that were “rescued” by the cohort features, i.e., correctly classified only by the classifier that had access to cohort information. **(E)** Mean (standard deviation) fraction of sibling embryos within a cohort (not including the transferred embryo/s) reaching blastulation was 0.5 (0.23) for negative cohorts versus 0.57 (0.17) for positive rescued embryos, Wilcoxon signed rank test on the differences from negative embryos mean found no statistical significance (p-value = 0.07). **(F)** Mean (standard deviation) cohort size was 10.64 (3.03) for negative cohorts versus 10.11 (3.39) for positive rescued embryos, Wilcoxon signed rank test on the differences from negative embryos mean found no statistical significance.

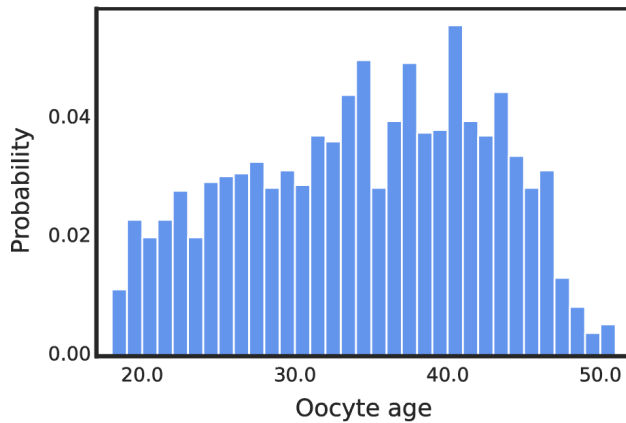


Figure S3: Distribution of oocyte age for all treatments. N= 2089. N cohorts = 1605. Mean (standard deviation) oocyte age was 33.99 (7.98).

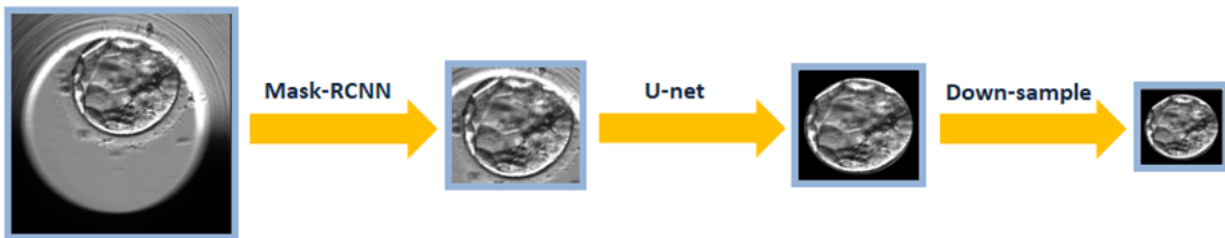


Figure S4: Oocyte segmentation pipeline (left-to-right). Mask-RCNN detects the embryo’s bounding box. The cropped image is segmented by a U-NET network. Finally the image is down sampled to 64x64.

תקציר

למידת מכונה המבוססת על תמונות עוברים מרגע ההפרייה ועד ההחזרה לרחם שמטרתה לשפר את בחירת העובר המוחזר ואת חיזוי ההסתברות של העובר להשתרש ברחם מחוללת מהפכה בתחום ההפרייה החוץ גופית (IVF). עם זאת, הרוב המכריע של העוברים לא מוחזר לרחם, וכיום מתעלמים מהמידע המגולם בעוברים אלה שתוצאות ההחזרה שלהם לא ידועות. במחקר זה אנו בודקים האם ובאיזו מידה המידע המקודד בעוברים "אחים" מאותו טיפול IVF תורם לביצועים של חיזוי השרשה מבוסס למידת מכונה. ראשית, אנו מראים כי תוצאת ההשרשה של העובר מתואמת עם מאפייני האחים בטיפול. שנית, אנחנו מדגימים כי המידע המקודד בעוברים אחים אלו משפר את ביצועי חיזוי ההשרשה של העובר המוחזר. שלישית, אנחנו מאפיינים אילו תכונות תרמו לחיזוי, כאשר אנחנו מתמקדים במיוחד בתכונות שהצילו עוברים בעלי תחזיות שגויות. התוצאות שלנו מצביעות על כך שמודלים לחיזוי השרשת עוברים יכולים להפיק תועלת מהדטה הרב, הזמין והלא מתוייג של עוברים אחים באותו הטיפול, על ידי הפחתת הרעש המובנה של העובר המוחזר.

אוניברסיטת בן-גוריון בנגב
הפקולטה למדעי ההנדסה
המחלקה להנדסת מערכות מידע

שימוש במידע לא מתויג מאחים באותו הטיפול כדי לשפר את התחזית להשרשה של עוברים בתהליך הפריית מבחנה

חיבור זה מהווה חלק מהדרישות לקבלת תואר מגיסטר בהנדסה

מאת: נעם צוקרמן

מנחה: ד"ר אסף זריצקי

תאריך: 7/11/2022

חתימת המחבר: נעם

תאריך: 7/11/2022

אישור המנחה/ים: אסף

תאריך:

אישור יו"ר ועדת תואר שני מחלקתית:

ספטמבר 2022

אלול תשפ"ב

אוניברסיטת בן-גוריון בנגב
הפקולטה למדעי ההנדסה
המחלקה להנדסת מערכות מידע

שימוש במידע לא מתויג מאחים באותו הטיפול כדי לשפר את התחזית להשרשה של עוברים בתהליך הפריית מבחנה

חיבור זה מהווה חלק מהדרישות לקבלת תואר מוגיטר בהנדסה

מאת: נעם צוקרמן

ספטמבר 2022

אלול תשפ"ב