Methods for predicting in vitro fertilization (IVF) embryo developmental potential by machine learning algorithms of video streaming data

Thesis submitted to the Senate of the Hebrew University of Jerusalem in partial fulfillment of the degree of "Doctor of Philosophy"

by

Itay Erlich

Supervised by Dr. Assaf Zaritsky

Submitted to the Senate of the Hebrew University of Jerusalem 08/2021

This work was carried out under the supervision of

Dr. Assaf Zaritsky

Acknowledgments

I would like to express my deep and sincere gratitude to Dr. Assaf Zaritsky at the Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva gave me the chance to conduct research, believed in me all the way, shared his profound knowledge of both biological and computer sciences aspects, and guided me throughout writing my doctorate. His patience, vision, sincerity and motivation inspired me deeply. Assaf – thank you so very much. It was a great privilege and honor to work and study under his tutelage.

I am grateful to Dr. Assaf Ben-Meir of the Obstetrics & Gynecology Fertility Department at the Hebrew University of Jerusalem for initiating this research and providing valuable knowledge in both biology and medicine throughout this research.

I would also like to express my gratitude to Prof. Lior Wolf, The School of Computer Science, Tel Aviv University, for accepting me for a joint research project, and giving me the opportunity to learn from his vast experience.

I also take this opportunity to express my gratitude to Dr. Roni Stern, Department of Software and Information Systems Engineering at Ben Gurion University of the Negev, for his help and support. I am grateful for his very valuable comments on this thesis. My special thanks go to Esther Singer for her professional work and all her efforts in editing this work.

Finally, I am so thankful to my dear wife and children for their love, understanding, and continuing support throughout this long process.

ABSTRACT

In vitro fertilization (IVF) is the most effective form of assisted reproductive technology. However, selecting the embryos in treatments with the best implantation potential remains a challenging task. Despite several decades of clinical practice, there is no reliable non-invasive method to identify the small fraction of embryos that possess the highest potential to develop into a blastocyst, which can then be implanted and hopefully proceed to term. Thus, to maintain reasonable pregnancy rates, current practice involves the transfer of multiple candidate embryos into the uterus. This results in clinical complications and health risks to the newborn and the mother, with many more multiple (embryo) pregnancies (over 10% of IVF pregnancies are twins or more).

To improve the pregnancy rates of IVF treatment, and avoid complications to both mother and fetus, only the embryos that are the most likely to implant should be identified and implanted. To this end, several methods have been proposed to rate embryos. Most methods so far have considered only 2D images of the embryo taken by a camera on top of a microscope moments before transferring.

The introduction of time-lapse incubation (TLI) systems in recent years provides continuous visualization of the embryos while maintaining them in optimal culture conditions. However, only a few studies have been conducted on the potential applications of time lapse imaging to assess embryos and identify the most viable ones. These are based on manual annotation, which involves a significant workload and human expertise, are restricted to later developmental stages and have only reported limited success.

This dissertation proposes an objective, automated and comprehensively evaluated system for grading models. First, I present a fully automated system that leverages deep learning to provide a continuous morphology-based embryo grading throughout its developmental process and demonstrate that it outperforms grading from a single snapshot. Second, I develop methods for grading embryos based on their morphokinetics - the timing of specific developmental stages - and comprehensively assess the interplay between morphology, morphokinetics and other clinical

factors in the prediction of embryo implantation potential. Third, I make the conceptual observation that current many machine learning based IVF approaches optimize a non-optimal problem because they combine the two tasks of implantation prediction and embryo ranking, and propose practical steps to overcome these limitations. Finally, I apply the insight gained by previous advances to the development of a pseudo contrastive labeled based embryo grading algorithm that allows learning from all untagged embryos and surpasses previous approaches and a group of eight senior professionals, in identifying viable embryos from non-viable embryos based on implantation results, developmental trajectory analysis, and genetic testing. Altogether, my dissertation provides methodologies that extend beyond the current state of the art as well as deeper understanding regarding developing and applying machine learning solutions for IVF. These contributions will hopefully drive the field forward and eventually provide better care for IVF patients.

1. INTRODUCTION

Infertility treatment by clinical in vitro fertilization (IVF) procedures involves the fertilization of an egg outside the human body. Upon fertilization, embryos are incubated in a culture medium for 2-7 days (Eskew & Jungheim, 2017). Once the embryos have reached the desired developmental stage, one or more embryos from the same cohort are transferred to the woman's uterus at the appropriate time in her menstrual cycle.

The clinical choice of the optimal time for transfer varies, and depends on maternal age, fertility status, and the number of fertilized oocytes. Preimplantation embryo development in mammals corresponds to 5 to 7 days of culture following oocyte fertilization and human embryos hatch on Day 6 (Chimote et al., 2016). There is a general consensus that the environmental conditions of embryo cultures in an incubator are not optimal (Mauri et al, 2001; Swain et al., 2016) and that most embryos are unable to implant in the uterus or produce a pregnancy (Jarvis, 2016). To avoid damaging the developmental potential of the embryos to implant or undermining the chances of pregnancy due to their prolonged culture in the incubator, the embryos need to be transferred as soon as possible.

This underscores the critical need to identify the embryos with the highest developmental potential. Since the ability to discriminate between high-quality and poor-quality embryos increases with time in culture, embryos tend to be transferred as late as possible. In practice, the compromise is to transfer embryos on Day 3, which corresponds to the 8-cell cleavage stage (Yang et al. 2015). In cases where there are a relatively large number of embryos from one woman, doctors are more likely to extend the time of culture of the embryos to five days and transfer the ones that have reached the blastocyst stage (Hatırnaz & Kanat Pektaş, 2017).

To date, procedures that can evaluate the developmental potential of cleavage-stage embryos, reported only a limited success (Alpha Scientists, 2011). Thus, more than one embryo is routinely transferred to obtain reasonable pregnancy rates. Along with increasing the pregnancy rates, multiple embryo transfers can also lead to multiple pregnancies (MP), which occur due to successful implantation of more than one embryo. MP rates depend on maternal age and typically

account for 10 to 15% of all IVF pregnancies of which >95% are twins (David et al., 2016). MPs are associated with a higher risk to the mother and the newborn (Khalaf, 2013). These complications place a significant clinical and financial burden on healthcare services worldwide and consume valuable resources. Eliminating this negative impact requires the transition from multiple- to single-embryo transfer (SET) schemes, which thus demand early-stage identification of the embryos with the highest developmental potential for implantation (Yang et al. 2015).

1.1. Microscopy and Algorithmic Techniques for Embryo Selection

Up to recent years, embryo selection during IVF was performed manually by clinical embryologists based on visual examination of morphological characteristics from embryos using an optical microscope at various stages of development such as the cleavage stage (Day 3, Holte et al. 2007) or even when the pronuclei are visible (Day 1, Mirroshandel et al. 2016), with limited success. To overcome this inability to accurately evaluate embryos at early developmental stages, culturing embryos to the blastocyst stage combined with grading protocols have frequently been implemented in both research and clinical practice (Blake et al., 2007; Gardner et al., 2000; Richter et al., 2001; Shapiro et al., 2008;; Papanikolaou et al., 2006; Papanikolaou 2008).One of the most common scoring systems used by embryologists is the Gardner's Score (Gardner et at. 2000), which evaluates embryos based on characteristics of morphological structures like inner cell mass (ICM), trophectoderm quality and embryo developmental advancement. Such manual grading is, however, extremely subjective, and there is a high level of variation between embryologists, leading to a wide range of variability between and within operators (Kragh et al. ,2019). As a result, standardized algorithms for grading and evaluating embryos are critical. Recent studies by Khosravi et al. (2019) and Kragh et al. (2019) showed high levels of accuracy in the classification of blastocyst features used by the Gardner Score. While this approach may be useful for standardizing embryo classification according to the Gardner score, it has limited potential to improve implantation rates over solutions offered more than two decades ago.

Other attempts have been made to replace manual morphological grading with automatic systems. As an initial step, hand-designed features were extracted to train traditional machine learning models such as SVM (Hassan et al. 2018, Chavez-Badiola et al. 2020), multivariate logistic

regression (Chen at al. 2016), genetic algorithm and decision trees (Guh et al. 2011), Bayesian networks (Corani et al. 2013) and neural networks (Chavez-Badiola et al. 2020). Other studies incorporated clinical data, such as patient/oocyte age, causes of infertility, stimulation of the oocyte, and semen analysis (Loendersloot et al. 2014; Raef et al. 2019). In spite of these advances, there is still no generalized method available to measure embryo development at all stages and times.

Unlike traditional microscopy that can only capture snapshots of a few discrete points in time, time-lapse incubation (TLI) allows for constant monitoring of fertilized embryos (Aparicio-Ruiz et al., 2018, Meseguer et al., 2011). Numerous publications have addressed various cleavage stage tasks such as predicting blastocyst formation, automatic annotations and abnormal genetic chromosomal detection (Chen et al., 2014, Milewski et al., 2011, Reignier, 2018, Swain, 2013, Malmsten et al. 2021). Several studies have demonstrated positive correlations between pronuclei markers and embryo viability. Specifically, markers such as the appearance of pronuclei and fading timing (tPNa, tPNf), the number of pronuclei, pronuclei shape, symmetry and joint path have been associated with blastocyst development and implantation (Aguilar et al. 2014; Lynette et al. 2000;Mingzhao et al. 2017). However, neither of these markers has ever been integrated into a scoring system, and not all embryos are placed early enough in the TLI to track the pronucleus stage. Other studies suggested various dynamic markers of embryo quality that can lead to new approaches to embryo selection (Basile et al., 2014, Hlinka et al., 2012; Meseguer et al., 2011; Milewski et al., 2016). The most common (patented) implantation scoring for cleavage stage embryos is the KID3-score, EmbryoScope's built-in algorithm (Vitrolife®), which gives a (one-time) prediction after 66 hours and consists of a 58% AUC (Adolfsson et al., 2018). Yet, as these features are acquired manually, it relies on each embryologist's experience, precision, and ability to recognize normal cleavage patterns from abnormal ones. As a result, intra-operator and inter-operator variability, across embryologists and different labs, are significantly higher than when using two-dimensional microscopy (Sundvall et al., 2013).

Due to technological advancements in cell culture media, blastocyst stage transfers have become increasingly popular. Compared to the cleavage stage, blastocyst embryos are better synced with the uterine milieu and contain more information on embryo viability (Glujovsky et al. 2016). However there is no clear evidence that blastocyst stage transfers are associated with higher rates

than cleavage stage transfers (Blake et al. 2007 , Glujovsky et al. 2016; Papanikolaou et al. 2006).

New markers and grading models for Day 5 blastocyst embryos have also been proposed based on TLI tracking. According to one study (Saraeva et al. 2019), spontaneous blastocyst collapse is correlated with poor pregnancy outcome. An earlier study (Kragh et al. 2019) demonstrated the use of TLI streaming to automatically grade the Inner Cell Mass (ICM) and Trophectoderm. The results indicated that TLI led to inferior results compared to a single microscopy image snapshot. In yet another study (Bori et al. 2020) different spatial parameters (mainly related to the Zona Pellucida(ZP), Inner Cell Mass (ICM) and Trophectoderm (TE) related) acquired at different development stages (mainly late blastocyst stages) were used as features to accurately predict implantation results. There is one major disadvantage of all these studies: they focus on a local task (e.g., ICM and TE ranking), or they use manually selected temporal and spatial features, which are sparse in the TLI stream, making them difficult to apply in practice because of the quantity of data and expertise needed. More importantly, they were all limited to late blastocyst embryos, whereas regular microscopy can also provide embryo scoring.

Overall, both cleavage stage and blastocyst stage algorithms ignore the extra information outside of a specific developmental stage, thus not using the majority of the collected images. To date, there is no algorithm that exploits TLI streaming as a whole; i.e., that translates continuous embryo streaming into continuous embryo scoring and can be applied to all embryos with no limitations (e.g., restriction to the cleavage stage, blastocyst stage, etc.). Similarly there is no algorithm that is fully automated and does not require any human intervention and can be applied at any time and developmental stage. Thus, current TLI based algorithms do not exploit the information that can be acquired from continuous tracking of the TLI, but rather draw on multiple images which are used individually.

1.2. Convolutional Neural Networks (CNN) for Embryo Selection

Deep Convolutional Neural Networks (CNNs) are deep learning algorithms that can take in an input image, assign importance to various aspects/objects in the image, and differentiate one from the other. In particular, CNNs can exploit both spatial and temporal information, possibly at the same time, using 3D convolutional layers or recurrent networks (Mikolov et al., 2010). CNNs have led to state-of-the-art results on a variety of problems such as image classification and segmentation, object detection, video processing, speech recognition and natural language processing (Frizzi et al., 2016; He et al. 2017; Larsson et al., 2016; Wahab et al, 2019). Inspired by how the human brain works, deep CNN uses multiple feature extraction stages that can automatically learn representations from the data. The growing availability of data and advances in computing power have accelerated research in CNNs.

A deep network framework called IVY, has recently been developed to determine implantation potential of blastocyst embryos from time-lapse videos (Tran et al., 2019). The objective of their study was implantation prediction. Analysis of their data, however, reveals they studied how to discriminate visually defective discarded embryos, a fairly easy classification task that resulted in an impressive AUC of 0.93.

1.3. My dissertation: automated and continuous methods for embryo grading based on TLI data

Studies based on TLI are often limited to later developmental stages, mainly Day 5 blastocysts, and require manual annotations of many spatial/temporal features from the TLI stream to obtain a single score.

To date, there is no accepted or standardized algorithm that takes advantage of the enormous amount of information acquired in the TLI and can be applied to all embryos in all developmental stages. Additionally, TLI has not yet been shown to outperform traditional microscopy-based scoring. As a result, despite the explosion in terms of information that the TLI provides, IVF treatments remain inefficient, costly, and can involve numerous health-related complications to the mother and the baby associated with multiple pregnancies.

Stated simply, although TLI provides continuous tracking of the embryo, embryo selection algorithms are only available at a given late time, typically only after 3-5 days after fertilization. In that sense, the efficiency of TLI is very low, and the huge streaming data are simply considered a convenient way to obtain multiple microscopy images without having to remove the embryo from the incubator. There is no holistic approach that translates this continuous streaming into continuous embryo evaluation. This is mainly due to the fact that the TLI stream has yet to be exploited as a whole, but rather is used to capture specific features of specific parts of the embryo. In turn, current TLI based methods cannot provide early indications of embryo viability or continuous monitoring which explains why the use of TLI at IVF is not widespread.

1.3.1. **Goals and Contributions of this Thesis**

The primary goal of this work was to develop novel, automated, accurate and continuous classifiers of embryo blastulation, implantation and viability based on the entire time-lapse video. The importance of these classifiers is that they exploit TLI streaming as a whole and can be applied to all embryos with no limitation of time or developmental stage. Further, they do not require any manual annotations which are time-consuming, and provide continuous estimations of embryo viability as early as Day 1, rather than a single score long after the oocyte was fertilized. Thus, unlike other algorithms, it can be use seamlessly in real-time IVF sessions.

As importantly, this thesis also aims to discover and compare different training schemes in order to better understand what information can be extracted by the TLI at different times. The dissertation also discusses a number of different aspects of training embryo grading models, including the optimization task, train and test data, and potential cofounders.

The main contributions of this thesis can be summarized as follows:

1. TLI's continuous monitoring is transformed into continuous embryo scoring, utilizing the entire TLI stream instead of just small fragments, resulting in the earliest scoring and best-in-class accuracy for both blastocyst formation and implantation predictions. Unlike existing approaches used in IVF treatments, none of the models here require manual annotations and/or spatial/temporal feature extraction. Mathematically, it provides an

embryo scoring function over time rather than simply a singular scalar embryo score (obtained long after fertilization time).

2. Analyze the types of information available in TLI data by comparing different AI models based on each type of information. We compare morphokinetic information which represents the global developmental path with morphological information which represents the most current embryo status and explain the findings in relation to current medical best practice. Furthermore, we demonstrate that a joint model using both types of information beats each of them in all conditions.
3. Examines the differences between embryo ranking and implantation prediction. Our study shows that while solving the former with the latter is common for choosing the right embryos to be transferred, it leads to a suboptimal outcome. Comparing models trained to predict implantation outcomes with and without a known oocyte age shows this suboptimality. Instead, we urge separating both subtasks in order to choose the best embryos for transfer.
4. Show the trade-off between learning from hard, yet noisy-labeled data versus more trivial, yet clean-labeled data, with detailed and compensating analysis.
5. Present a pseudo contrastive label learning framework in the presence of noisy and poorly known data, that significantly outperforms fully supervised learning scheme.

1.4. **Organization**

This dissertation includes four projects, each packaged as an independent manuscript and in different stages of publication. Each project constitutes a chapter as follows:

The first study presents deep learning classifiers of blastocyst formation and implantation based on short-time (termed packets) morphological features. We were inspired by the idea that embryo grading may be achieved by finding unseen biomarkers within the TLI data that can be used to classify the embryos in some sense (implantation probability, blastocyst formation, viability, etc.). A CNN framework is presented to reveal unseen biomarkers seamlessly. It incorporates a time-dependent architecture with shared weights between times, an adaptive weighted hinge loss where the scoring for several packets is optimized to ensure correct embryo classification, and a multi-

frame (non-linear) fusion of the single frame scoring into a single embryo prediction scoring method. This research was extended and published as a separate study (Kan-Tor et al., 2020).

The second study introduces two fully automated morphokinetic and morphology models for predicting embryo viability and compares them. It then integrates morphology with morphokinetics and shows that the combination of both improves embryo classification accuracy over both the morphology and the morphokinetic models at all times. This study was performed with Dr. Assaf Ben-Meir and is currently under review in Human Reproduction.

The third study discusses the practical task of choosing embryos for transfer from a cohort of siblings' embryos. In particular, the difference between the task of predicting implantation and grading embryos by their viability is discussed and the impact of oocyte age on both tasks is analyzed. It is shown that although learning in the presence of known oocyte age improves implantation prediction, it lessens the ability to choose the most viable embryo from a cohort of same aged embryos, and thus should not be used to pick the best embryo for transfer from a cohort of related embryos. Instead, the practical scheme should consist of 2 models. The first ranks embryos based purely on embryonic related information, such as visual morphology and morphokinetics. Then, a second model determines the potential of each embryo to progress a full term, considering maternal information, to decide how many embryos should be collectively transferred. This study was submitted.

The forth study introduces a pseudo-contrastive-label learning framework to grade embryos by their viability. Approximately only 15% of all embryos have a known implantation outcome. The remainder are embryos are discarded, frozen or jointly transferred (Tran et al. 2019). Further, the non-implanted embryos have noisy labels, because a negative outcome may be due to maternal or other non-embryonic related reasons. Therefore, a learning framework that considers the noise in the negative (termed KIDn) label would likely lead to better results. In this chapter we propose a Pseudo Contrastive Labeling scheme that makes it possible to use any embryo, regardless of its original label (i.e., positive (termed KID-P), KIDn, ambiguous (termed KIDuk) or Discarded). As a result, the training size increases six fold since all the embryos are added to the training set, rather than just the fully labeled ones. We demonstrate the efficiency of the new scheme over learning

based on the known implantation labels. This study was performed with Prof. Lior Wolf and is currently under review in Scientific Reports.

Developing automated deep-learning classifiers for predicting embryo blastulation and implantation at early cleavage stages

Itay Erlich,[1*,**] Yoav Kan-Tor,[1] Tamar Amitai,[1] Yuval Or,[4] Zeev Shoham,[4] Arieh Horowitz,[3] Iris Har-Vardi, [5] Matan Gavish,[1] Assaf Ben-Meir[3] and Amnon Buxboim[1,2]

[1]*The Alexender Grass Center for Bioengineering, School of computer Science and Engineering, Hebrew University of Jerusalem, Israel*

[2]*The Institute of life sciences, Hebrew University of Jerusalem, Israel*

[3]*Fertility and IVF Unit, Department of Obstetrics and Gynecology, Hebrew University Hadassah Medical Center, Israel*

[4]*Fertility and IVF Unit, Kaplan Medical Center, Rehovot, Israel, Hebrew University, Jerusalem, Israel.*

[5]*Fertility and IVF Unit, Department of Obstetrics and Gynecology, Soroka University Medical Center, Ben-Gurion University of the Negev, Beer Sheva, Israel*

** email: ityer82@gmail.coml*

**** unpublished

Summary

In IVF treatments, early identification of embryos with high implantation potential is required for avoiding clinical complications associated with multiple-embryo pregnancy. Current classification tools evaluate embryo potential based on manual annotation of morphological features and morphokinetic events. Despite the availability of time-lapse incubators that record preimplantation development, there is no reliable non-invasive method for evaluating embryo potential. We employ deep learning tools to develop automated, accurate and standardized classifiers of embryo blastulation and implantation based on time-lapse videos. Our learning set consists of over 12,000 embryo video files and clinical metadata obtained from five medical centers. Our prediction of embryo blastulation and implantation outperforms current state-of-the-art classifiers. Moreover, our classifiers spontaneously tuned to times of high information entropy based on cleavage-states distributions, providing insight into learning mechanisms.

Clinical implementation of our classifiers will improve IVF treatments by shortening time-to-pregnancy and facilitating single embryo transfers.

Introduction

Today, in vitro fertilization – embryo transfer (IVF-ET) procedures account for 1.7% of all newborns in the United States.[1] After 40 years of clinical practice, there is no reliable non-invasive method available for identifying the small fraction of embryos that possess the highest potential to develop into a blastocyst, to implant and to proceed to full term labor.[2,3] To maintain reasonable pregnancy rates, current practice involves the transfer of multiple non-genetically diagnosed embryos into the uterus. As a result, clinical complications and health risks to the newborn and the mother which are associated with multiple (embryo) pregnancy cannot be avoided.[4]

The introduction of time-lapse incubation systems provided continuous visualization of the embryos while maintaining them in optimal culture conditions.[5,6] These video recordings facilitated the design of classification algorithms that evaluate embryonic potential based on morphokinetic annotation of discrete events (time points of appearance and fading of the pronuclei and cleavage events).[7,8] Notably, morphokinetic classification may limit accurate assessment of embryo quality, which can just as well be encoded within other visible dynamic features.[8] In addition, manual annotation requires a significant workload and human expertise, and is prone to variations between annotators.

To overcome these limitations, embryos can be evaluated using automated computer-based artificial intelligence tools. Convolutional neural networks (CNN's) are among the most impactful innovations in computer vision in recent years. CNN's facilitated automated object detection, speech recognition, face recognition and medical image segmentation [12-19] that made significant advances in biomedical imaging, security and biometric applications, and autonomous vehicles.[10-11]

Here we report the development of automated, accurate and standardized classifiers for identification of the most competent embryos at early preimplantation stages. Since deep

learning requires large datasets, we generated an expansive database, termed SHIFRA, that includes video files of tens of thousands of embryos and their transfer and implantation rates that were collected from five hospitals across Israel. We combine classification algorithms of embryo blastulation and embryo implantation, which demonstrate superior prediction compared with KIDScore-D3$^{LTD\ 9}$. Mechanistic analysis reveals a relationship between blastomere cleavage states and the visual encoding of embryo developmental competence. Deep learning automated classification holds the potential to improve clinical outcomes by facilitating single embryo transfers while maintaining high implantation rates.

Results

The SHIFRA database: video files and clinical metadata for thousands of embryos

The implementation of time-lapse incubators in IVF clinics provided for the first time a large dataset of video files of high optical quality that offer detailed dynamic description of embryo development in association with medical background and clinical outcome.[20] We constructed the SHIFRA database, which includes information on tens of thousands of embryos obtained from nine incubators (Embryoscope time-lapse system, Vitrolife) located in five medical centers all across Israel (Fig. 1a, Table S1). Each embryo is marked by maternal characteristics (maternal age), embryonic fate (fresh-transfer/frozen/discarded) and outcome statistics (single/multiple embryos transfer and implantation rates). Implantation in the uterus was determined based on established protocols as determined by the number of gestational sacks and heart beats measured on week five and six of pregnancy. In the case that the number of transferred and implanted embryos is equal, the embryos are labelled Known Implantation Data-positive (KID_p). KID-negative (KID_n) corresponds to the case of no implanted embryos, and KID-unknown (KID_u) labels embryos whose implantation outcome cannot be determined (e.g., three transferred embryos, one or two implanted embryos). Similarly, embryos that were incubated longer than three days and show start-of-blastulation (SB) were classified BLAST-positive ($BLAST_p$). Since BLAST-negative ($BLAST_n$) embryos cannot be identified by direct observation, we applied a statistical algorithm that identifies the embryos which were arrested at earlier developmental stages and cannot reach blastulation. A detailed description of this protocol, which is based on the time windows that are associated with each morphokinetic state and their intervals, is provided in the Methods section (Fig. S2).

Time-lapse videos record up to six days of preimplantation development with 15-to-20 min intervals. At each time point, the entire three-dimensional (3D) volume of the embryos is captured by recording a seven-focal plane z-stack ranging from -45 to +45 μm (Fig. 1b-ii, 500 x 500 pixels, 8 bit grayscale). Based on these time-lapse imaging, morphokinetic annotation of the embryos in the database was performed by expert and trained embryologists in accordance with established protocols[8] and validated by performing quality assurance (QA, Fig. S1a-b). Based on a total of 12,102 embryos, we calculated the temporal distributions of morphokinetic events (Fig.

1a, Table S2) and their consecutive intervals (Fig. 1b, Table S3). Morphokinetic distributions clearly show the separation between first (tPNf-to-t2), second (t3-to-t4) and third (t5- to-t8) cleavage cycles where the transitions between cell cycles typically lasts twelve hours (Fig. 1b). The small fraction of embryos that show simultaneous cleavage events (t2 and t3; t4 and t5) are associated with direct unequal cleavage[7].

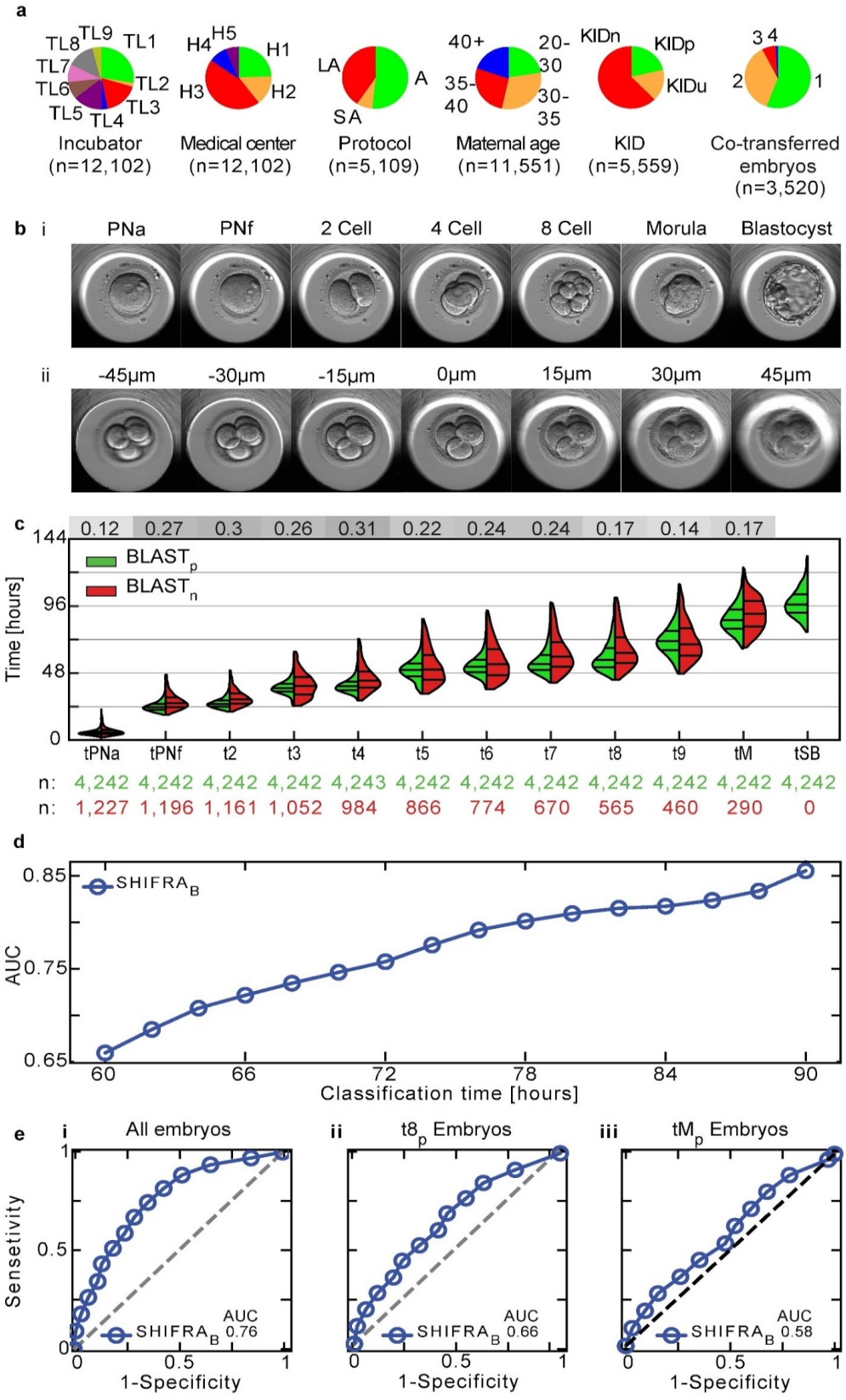

a
TL8
TL9
TL1
TL7
TL6
TL2
TL5
TL4
TL3
Incubator
(n=12,102)
H4
H5
H1
H3
H2
Medical center
(n=12,102)
LA
A
SA
Protocol
(n=5,109)
40+
20-30
35-40
30-35
Maternal age
(n=11,551)
KIDn
KIDp
KIDu
KID
(n=5,559)
3 4
2
1
Co-transferred embryos
(n=3,520)
b i
PNa
PNf
2 Cell
4 Cell
8 Cell
Morula
Blastocyst
ii
-45µm
-30µm
-15µm
0µm
15µm
30µm
45µm
c
0.12 0.27 0.3 0.26 0.31 0.22 0.24 0.24 0.17 0.14 0.17
144
96
48
0
Time [hours]
BLASTp
BLASTn
tPNa tPNf t2 t3 t4 t5 t6 t7 t8 t9 tM tSB
n: 4,242 4,242 4,242 4,242 4,243 4,242 4,242 4,242 4,242 4,242 4,242 4,242
n: 1,227 1,196 1,161 1,052 984 866 774 670 565 460 290 0
d
0.85
0.75
0.65
AUC
SHIFRAB
60 66 72 78 84 90
Classification time [hours]
e
i All embryos
ii t8p Embryos
iii tMp Embryos
Sensetivity
1
0.5
0
SHIFRAB AUC 0.76
SHIFRAB AUC 0.66
SHIFRAB AUC 0.58
0 0.5 1
1-Specificity

Figure 1 | Deep learning classifier of embryo blastulation trained on an expansive database of preimplantation development. (a) The SHIFRA database (>12,000 embryos in total) includes clinical (Time-lapse incubator; Medical center; superovulation protocol), maternal information (maternal age), number of co-transferred embryos (3520 transfer cycles) and implantation rates. Abbreviations: A- Agonist; SA-Short Agonist; LA-Long Agonist. (b) Preimplantation development of embryos that are cultured inside a time-lapse incubator are recorded at 15-to-20 min intervals over up to six days. (i) A series of representative snapshots show embryonic developmental stages. (ii) A z-stack of seven focal planes, separated by 15 μm, are recorded at each time point. A z-stack of the 4-cells stage is shown. (c) Morphokinetic distributions of BLASTp and BLASTn embryos show significant overlap as evident by low KS test scores (top row). (d) BLAST prediction performances by $SHIFRA_B$, measured using area under the curve (AUC), increases monotonically as a function of time of classification. (e) ROC curves at 72 hours for (i) all embryos, (ii) high quality t8-positive embryos (embryos that reached 8C stage), and (iii) very high quality tM-positive embryos (embryos that reached Morula stage).

Deep learning of embryonic developmental potential

Examination of the morphokinetic distributions of BLAST-labeled embryos reveals almost complete overlap, indicating that BLAST classification is a complex task that cannot be resolved based on average morphokinetic profiles per se (Fig. 1c, KS test: 0.1 to 0.3). This conclusion is further supported by comparing the morphokinetic distributions of KID-labeled embryos (Fig. 2a; 0.1 < KS test < 0.2). Therefore, we employed advanced deep learning tools to develop an automated, accurate and standardized classifier of embryo blastulation and embryo implantation. Following strict methodologies, BLAST and KID validation sets were generated by randomly removing 18% of the positive and negative labeled embryos. Validation sets were maintained uncontaminated until prediction by BLAST and KID classifiers was performed after training was completed.

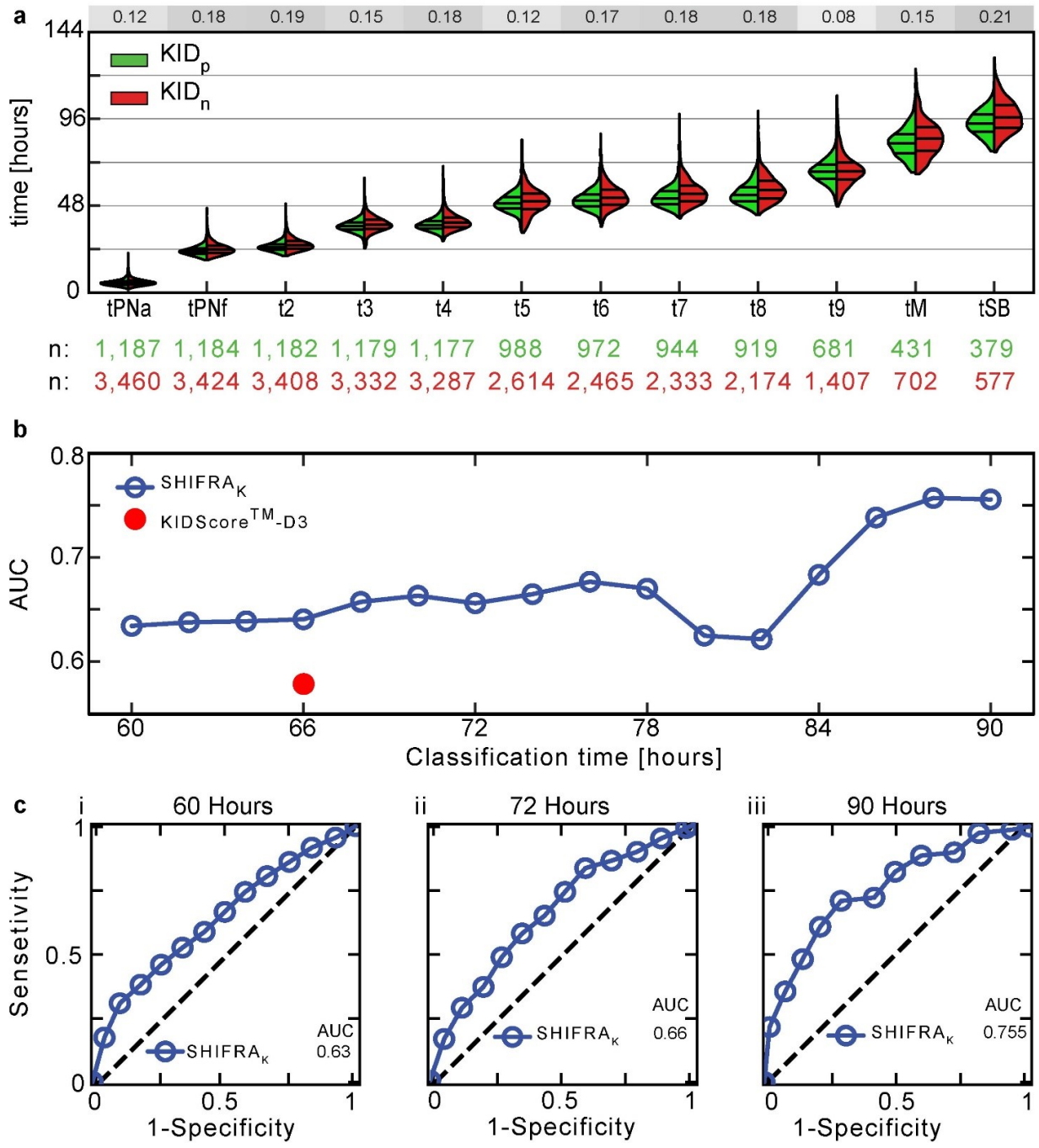


Figure 2 | SHIFRA$_K$ is an accurate, automated and standardized classifier of embryo implantation. (a) Morphokinetic distributions of KID$_p$ versus KID$_n$ embryos show significant overlap as evident by low KS test scores (top row). (b) Embryo implantation prediction performances by SHIFRA$_K$, measured by the AUC, increase as a function of time of classification. The drop in AUC is the result of a decrease in the number of transferred embryos for training with incubation times longer than 80 hours (Fig. S3a-b). Compared with KIDScore-D3 (Vitrolife, purple mark), SHIFRA$_K$ demonstrates superior classification. (c) ROC curves of implantation prediction at (i) 60 hours compared, (ii) 72 hours and (iii) 90 hours.

In general, the convergence of supervised neural networks depends on both the size of labeled objects and its variability. We performed preprocessing of the input video files, ~100 MB, as follows (Fig. 3a): (1) Identification and discarding of empty well images (Fig. S4a, see Methods). (2) Cropping of square ROI's that contain the embryos (Fig. S4b, see Methods). (3) Down-sampling of cropped embryo ROI's. Next, five consecutive preprocessed frames of the same focal plane and the same embryo were grouped into packets. In this manner, we captured the dynamic nature of preimplantation processes that can last 60 to 90 minutes (e.g. cleavage). Within each training batch, each embryo was thus represented by four packets, which is equivalent to a decrease of the raw video file to less than 1 MB.

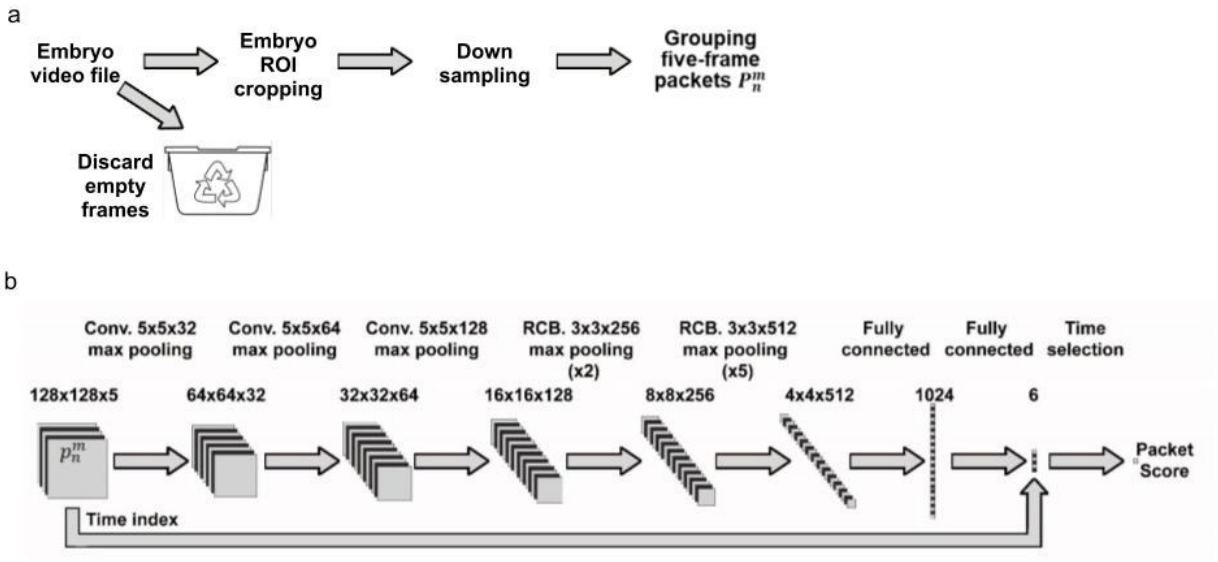


Figure 3| Neural network design. (a) Preprocessing of the video files of the embryos into a small set of five-frames packets, P_n^m, accounts for a ~150 fold dimension reduction. Packets are defined by their embryo index, m, and their time index, n. (b) Network design is sketched. Conv: Convolution. RCB : Residual convolution block. Time: Incubation time of the middle frame.

CNN design consisted of 13 layers as illustrated in figure 3b. The architecture of the network is described in details in the Methods section. To emphasize preimplantation dynamics, packets were scored relative only to other packets of the same 12-hours' time windows. We postulate

that embryo developmental potential is marked by scarce events. Namely, a small number of 'positive' packets are sufficient to score a positive embryo whereas a negative embryo must be scored negative by most packets. This principle is implemented by modifying the soft hinge loss to optimize the performances of the classifier (see Methods). CNN training by BLAST-labeled and by KID-labeled embryos converged following ~three epochs. Following training, the network was applied to BLAST-labeled and to KID-labeled embryos to evaluate their respective developmental potentials as described below. To obtain as much information as possible, each embryo was scored using all of its packets up to the time of classification. We found that classification performances were not compromised when we included only the packets that belong to the twelve hours that precede the time of classification and ignore earlier ones (Methods).

Early prediction of embryo blastulation

The capacity of an embryo to reach the blastocyst stage inside the incubator marks its developmental competence and is linked with the capacity of the embryo to implant in the uterus[3]. Hence, early identification of $BLAST_p$ embryos is of high clinical value. We developed the $SHIFRA_B$ classifier by performing deep learning of BLAST-labeled embryos. Our first aim was to elucidate the relationship between classification time and BLAST-prediction. For each classification time, only embryos with video recordings longer than that were considered for training, where only frames earlier than the classification time were used. Overfitting of $SHIFRA_B$ was invalidated by the overlapping ROC curves of BLAST-classification of the train, test and validation sets (Fig. S5a). BLAST-prediction, measured by the area under the ROC curve (AUC), increased monotonically from 0.66 at 60 hours and 0.76 at 72 hours to 0.85 at 90 hours (Fig. 1d). As expected, BLAST classification becomes more difficult when applied to cohorts of high-quality embryos that either reached 8-cells stage ($8C_p$) or reached Morula stage ($MORULA_p$) as compared with all embryos (Fig. 1e). Nevertheless, BLAST prediction remained well above random, thus demonstrating the classification robustness of $SHIFRA_B$.

Positive and negative predictive values (PPV and NPV) were calculated as a function of $SHIFRA_B$'s threshold value at 72 hours (Fig. S5b). Generating a binary classifier based on $SHIFRA_B$ requires

choosing a threshold value, which should be directed by clinical considerations. Clinical implementation of single embryo transfers (SET's) requires minimizing false-positives. We therefore plot the confusion matrix at threshold value 41, which satisfies PPV=0.9 and NPV=0.33 (Fig. S5c). Taken together, $SHIFRA_B$ demonstrates high classification performances with important clinical value, thus serving as a proof of concept of automated embryo classification.

Early prediction of embryo implantation

$SHIFRA_B$ demonstrates a proof-of-concept of fully-automated classification of embryo developmental competence. Unlike blastulation, which depends on the quality of the embryo, implantation also depends on the receptivity of the uterus and on other factors that are not taken into account in the learning process. Hence, we designed $SHIFRA_K$ by integrating six neural networks in parallel to avoid overfitting. The design and training of each network was identical to $SHIFRA_B$ as described above. Three neural networks were trained on BLAST-labeled embryos and three networks were trained on KID-labeled embryos. Importantly, $SHIFRA_K$ is a KID-classifier where all six networks predict implantation of KID-labeled embryos. As a result, six scores are generated for each embryo. A final implantation score was obtained by performing weighted summation using linear soft support vector machine (SVM, see Methods) with respect to the KID-label of each embryo.

KID classification performances were evaluated by applying $SHIFRA_K$ on the validation set. At each classification time, a ROC curve was generated and the AUC was calculated (Fig. 2b). The AUC changes non-monotonically with classification time due to the combined opposite effects of the overall gain of information during embryo incubation and the decrease in the number of KID-labeled embryos that are available for training between 60 and 90 hours (Fig. S3). AUC was 0.63 at 60 hours, 0.66 at 72 hours and 0.75 at 90 hours (Fig. 2c). The improvement in KID classification between 72 and 90 hours owes to an increase in PPV where NPV remains almost unchanged (Fig. S6a-b). KID prediction is illustrated at 72 hours and at 90 hours by choosing threshold values and applying $SHIFRA_K$ on validation set embryos (Fig. S6a-ii, S6b-ii). Finally, we compared KID-classification of $SHIFRA_K$ with KIDScore-D3[9]. Both classifiers were applied on the same uncontaminated validation set embryos at 66 hours from fertilization. Satisfyingly, $SHIFRA_K$

generated a better prediction of embryo implantation (AUC 0.65 versus 0.59), which was automated and independent of morphokinetic events.

Classification by SHIFRA$_B$ and SHIFRA$_K$ is robust to differences in maternal age

Throughout the pre-menopause life span of the woman, oocytes are continuously exposed to various stress mediators[21]. Hence, maternal age is an important determinant of female fertility[22]. Based on the SHIFRA database, we performed a broad comparison between the morphokinetic statistics of young (age<32) versus aged (age>38) women (Fig. 4a). Based on our database, which is over-represented by couples that experienced conception difficulties, the differences between young and aged morphokinetic distributions appear to be negligible (Kolmogorov-Smirnov test score < 0.06). Moreover, median morphokientic values of young embryos appear to exceed aged embryos only starting from the Morula and Start-of-Blastulation events (Fig. 4a). Consistent with the overlap in morphokinetic distributions, the fraction of BLAST$_p$ embryos out of all BLAST-labeled embryos is independent of maternal age (Fig. 4b-i), yet the fraction of positively implanted embryos out of all transferred embryos is four-fold higher in young women (Fig. 4b-ii). These differences in implantation potential but not in blastulation potential indicate that the oocytes of young and aged women that undergo infertility treatments possess the same developmental quality but the receptivity of the uterus is compromised in aged women. Finally, we compare between the ROC curves of SHIFRA$_B$ and SHIFRA$_K$ at 72 hours and at 90 hours (Fig. 4c-d). The AUC's of young and aged embryos were comparable with the prediction of the entire pool of BLAST-labeled (Fig. 1e-i) and of KID-labeled (Fig. 5c-ii) embryos. This excludes the option that SHIFRA$_B$ and SHIFRA$_K$ are skewed towards young or aged embryos. It is thus demonstrated that both SHIFRA$_B$ and SHIFRA$_K$ are robust to differences in maternal age.

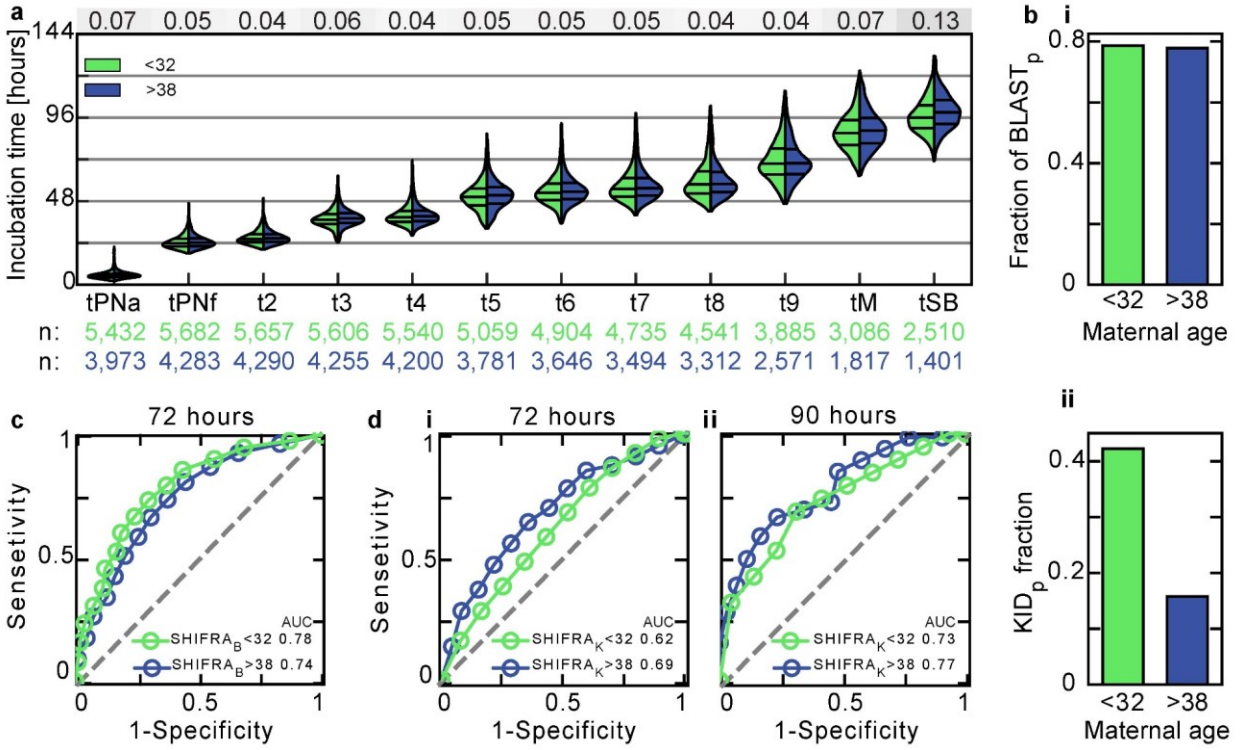


Figure 4| $SHIFRA_B$ and $SHIFRA_K$ classifiers are robust to differences in maternal age. (a) Morphokinetic comparison between young (age<35) and aged (age>38) embryos show negligible differences (Kolmogorov-Smirnof test score < 0.1). The fraction of embryos that reached blastulation (b-i) is very similar between the two age groups (0.78 and 0.77 accordingly) non the less (b-ii), the fraction of KIDP is scientifically higher for embryos with <35 maternal age (0.40 and 0.1 accordingly). ROC curves of (c) BLAST prediction by $SHIFRA_B$ and (d) KID prediction by $SHIFRA_K$ are shown for young (age<35) and aged (age>38) women at , (i) 72 hours and (ii) 90 hours.

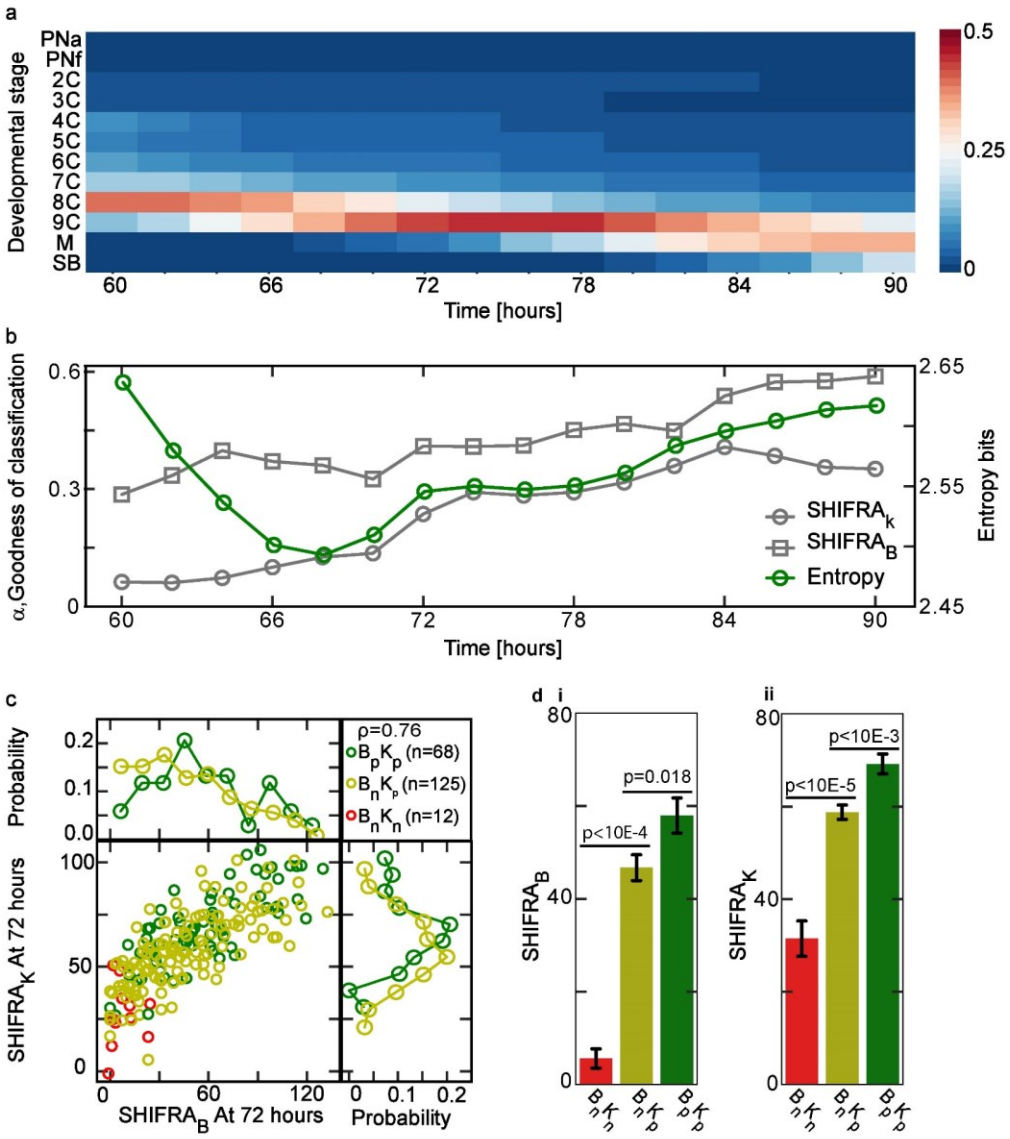


Figure 5 | BLAST and KID deep learning classifiers share common features correlated with morphokinetic information entropy. (a) Each column shows the morphokinetic state distribution of embryos at specific time points ranging between 60 and 90 hours from ICSI. Probability of morphokinetic events are color coded. Most probable state propagates from 8-cells (60 to 66 hours) to 9-cells (68 to 84 hours) and to Morula (86 to 90 hours). (b) Information entropy (green symbols) is calculated based on the morphokinetic state distributions (a). A score of the goodness of classification, γ, is plotted for $SHIFRA_K$ and $SHIFRA_B$ (gray symbols, see Methods). From 66 hours onward, the information entropy and the goodness of classification contours share a stepwise trend increasing at 72 and at 84 hours. (c) Scores obtained by $SHIFRA_K$ are plotted as a function of $SHIFRA_B$ for embryos that are BLAST and KID co-labeled. Top and right panels show the corresponding histograms, respectively. $BLAST_{n/p}KID_{n/p}$ embryos are abbreviated by $B_{n/p}K_{n/p}$. (d) Average $SHIFRA_B$ (i) and $SHIFRA_K$ (ii) scores are shown. B_nK_n embryos are scored lowest and B_pK_p embryos are scored highest by ***both*** classifiers.

Direct unequal features

Training of the BLAST and KID classifiers was performed with no specific reference to direct unequal cleavage (DUC). Compared with other DUC embryos, DUC1 (namely $t_2 = t_3$, Fig. S2d) is responsible for the most significant decrease in embryo blastulation, implantation and euploid rate, and holds the most far reaching clinical implications.[23] Therefore, we focus here on the classification outcome of DUC1 embryos (Fig. S7a). In the SHIFRA database, there are 580 DUC1 embryos and 11522 non-DUC1 embryos. Since DUC1 embryos were deselected for transfer, only 119 DUC-1 embryos were transferred compared with 5652 non-DUC1 transferred embryos out of which 17 DUC-1 embryos and 1745 non-DUC1 embryos successfully implanted (KID_P and KID_u combined). Therefore, a crude estimation indicates that the likelihood of implantation of DUC1 embryos is 2.2- lower. Consistent with this low implantation potential, we found that both KID_n and KID_p DUC1 embryos obtain low scores by $SHIFRA_K$ compared with non-DUC1 embryos (Fig. S7b). Evidently, by scoring DUC1-embryos low, $SHIFRA_K$ is effectively deselecting these embryos. We note that prediction by $SHIFRA_K$ takes into account frames acquired during the 12-hours' that precede time of classification (60 to 72 hours). Namely, images of first and second cleavage were not included in KID prediction. Hence, DUC1 embryos are deselected by $SHIFRA_K$ based on visual information that propagated forward by at least two days.

How do $SHIFRA_B$ and $SHIFRA_K$ work?

Obtaining insights into how deep learning classifiers function and uncovering the determinant visual features is a difficult task. We therefore focused on identifying the time windows that are used by the classifiers to predict blastulation and implantation. The probability distributions of finding the embryos at specific embryonic states as a function of time is summarized in figure 5a. These distributions are complementary to the statistics of morphokinetic events, which correspond to the transition between these states (Fig. 5a). For example, at 72 hours most of the embryos are found at the 8-cells state whereas morula is the most probable state at 90 hours. To quantify the amount of information encoded in each two-hours' time window, we calculate the

information entropy based on the probability distributions of the measured embryonic states (Fig. 5b). In the case that all embryos are found at a specific state, no information is provided by this time window. Consistently, the entropy is zero. Once a few embryos propagate to different states, entropy will increase. Maximal entropy is obtained when embryos are equally distributed across all possible states. We find that the entropy decreases between 60 and 68 hours and increases in a stepwise manner from 68 to 90 hours (Fig. 5b, green line). To obtain insight into how visual markers of embryo potential are encoded, we plot the so called goodness of classification, $\alpha(t)$, for both classifiers (Fig. 5b). $\alpha(t)$ increases in the case that the classifier gave a high score to packets of a positive-labeled embryos and low scores to packets of negative-labeled embryos and vice versa (see Methods). Remarkably, we find that the information entropy and the goodness of classification of both $SHIFRA_B$ and $SHIFRA_K$ show an overlapping trend (from 66 hours forward), which increases in a stepwise manner at 70-to-72 hours and at 82-to-84 hours from fertilization (Fig. 5b).

To estimate the degree of classification overlap between $SHIFRA_B$ and $SHIFRA_K$, we considered the embryos that were both BLAST and KID labeled and plotted their classification scores as obtained by $SHIFRA_K$ as a function of $SHIFRA_B$ (Fig. 5c). The positive correlation between $SHIFRA_B$ and $SHIFRA_K$ (Pearson correlation coefficient 0.76) is indicative of common visual features that are shared by both classifiers. Importantly, $BLAST_n$-KID_n (B_nK_n) embryos were scored lower than $BLAST_p$-KID_n (B_pK_n) embryos not only by $SHIFRA_B$ but also by $SHIFRA_K$ despite the fact that all these embryos failed to implant (Fig. 5d-i). Consistently, B_pK_n embryos were scored lower than B_pK_p not only by $SHIFRA_K$ but also by $SHIFRA_B$ despite the fact that all these embryos successfully initiated blastulation (Fig. 5d-ii). Hence, $SHIFRA_B$ is sensitive to implantation potential and $SHIFRA_K$ is sensitive to blastulaton potential. Taken together, we conclude that visual features are detected and shared by $SHIFRA_B$ and $SHIFRA_K$. The sensitivity of $SHIFRA_K$ to BLAST-labeling cannot be attributed to the three subnetworks that had been trained on BLAST-labeled embryos since all subnetworks were integrated using an SVM that was trained only on KID-labeled embryos. The fact that visual features that are associated with blastulation contribute to prediction of implantation and vice versa suggests that an overall embryonic developmental quality is encoded

by specific visual elements at specific time points, consistent with the overlapping dynamic information trends exhibited by $SHIFRA_B$ and $SHIFRA_K$ discussed above (Fig 5b).

Discussion

Automated, standardized and accurate classification of embryo blastulation and implantation

Implementing single-embryo transfer methodology holds the potential of reducing clinical complications but maintaining reasonable implantation rates requires means of selecting the highest-potential embryos for transfer.[25] With respect to a dozen or less embryos typically obtained from individual couples, the decision-making process that the physicians are required to carry out is complex. Each embryo can be either transferred, frozen, resume incubation or discarded. Since the goal is reaching live birth from any embryo, a comprehensive strategy is required that includes multiple transfer cycles. Here, we address the first step – scoring the developmental potential of individual embryos. We report the development of the first fully-automated classifiers of embryo developmental competence that are based on deep learning of video recordings. The automation of embryo selection offers several advantages with important clinical implications. (1) Standardization: $SHIFRA_B$ and $SHIFRA_K$ provide a standardized platform that is invariant across IVF clinics and is independent of human decision making, emotional state, and fatigue. (2) Quick and dynamic: Computer-based automated embryo classification is quick and can offer a dynamic profile of the developmental potential during incubation. (3) Accuracy: $SHIFRA_K$ demonstrates higher accuracy compared with KIDScore-D3 (Fig. 2b). We find that removing embryos from the training set leads to a decrease in classification performances (data not shown), indicating that increasing the number of labeled embryos will further improve classification accuracy and robustness. (4) Reducing workload: The clinical staff (specifically the embryologists) are often required to dedicate a significant time for embryo evaluation. In this regard, $SHIFRA_B$ and $SHIFRA_K$ can serve as a computer-based substitute.

How is the information about embryo quality encoded in time-lapse imaging?

Morphokinetic-based classification of embryo blastualtion[26] and implantation[7,20] confirm that the developmental potential of the embryo can be encoded, at least partially, in its temporal profile of morphokinetic events. As morphokientic-based classifiers are limited to a discrete representation of the embryos, unbiased embryo classification by direct learning of time-lapse imaging raises the question of whether additional developmental information is contained within

other visual dynamic or static features. We address this question using a twofold approach. First, we uncover the time windows that are associated with correct classification of positive and negative labeled embryos (Fig. 7b). We found that embryo classification both by $SHIFRA_B$ and by $SHIFRA_K$ showed a stepwise trend. Classification accuracy is improved when approaching 72 hours, which is associated with embryonic genome activation (EGA)[26], and when approaching 84 hours. An insight into the nature of this stepwise trend is obtained by comparison with the information entropy kinetics. Remarkably, the information entropy reproduces this stepwise trend (from 66 hours onward), suggesting that $SHIFRA_B$ and by $SHIFRA_K$ are sensitive to embryo cleavage events (Fig. 7b). Secondly, we considered the group of embryos that were labeled both in terms of blastulation and implantation (Fig. 7c). Since the scores obtained by $SHIFRA_B$ are higher for KID_P embryos and the scores obtained by $SHIFRA_K$ are higher for $BLAST_P$ embryos, we conclude that BLAST classification and KID classification share the same visual dynamic features (Fig. 7d). Taken together, our results suggest that deep learning of BLAST and of KID labeled embryos is sensitive to visual dynamic features that are linked with embryo cleavage events and are shared by both classifiers.

Clinical implications

Since deep learning requires a large number of labeled embryos that cannot be provided by prospective clinical experiments, we used retrospectively obtained data. In these IVF treatments, embryos were preselected for transfer based on morphological and morphokinetic criteria. As a result, KID_p, KID_n and KID_u embryos share visual characteristics that are not representative of the entire embryo population. Evaluating the clinical contributions of the SHIFRA classifiers thus requires a prospective clinical experiment.

Based on this work, we draw the following clinical implications and discuss two approaches for early and late embryo transfers. Without reliable non-invasive classification tools, day-5 transfers were introduced in order to screen embryos that fail to reach blastulation. $SHIFRA_B$ and $SHIFRA_K$ provide alternative means of reaching day-3 SET. In this manner, we prevent prolonged culture periods that were implicated with a decrease of embryonic potential as well as avoid risks associated with cryopreservation of day-5 expanded blastocysts. We propose combining $SHIFRA_B$

and $SHIFRA_K$ to identify the embryos with the highest developmental potential on day-3. In case there are embryos that are scored high, transfer of the embryo with the highest score should be considered. The remaining high-score embryos can then be frozen. Low-scored embryos should be further cultured, thus enabling the embryologists to discard the ones that become arrested prior to blastulation. A second option that this work supports is to perform early day-4 transfers (90 hours from ICSI). Relative to 72 hours, we find that the classification performances of $SHIFRA_B$ (AUC 0.76 versus 0.86) and of $SHIFRA_K$ (AUC 0.66 versus 0.76) improve significantly. Since blastocyst expansion occurs later, the improvement in identifying high-quality embryos at 90 hours does not introduce the reported difficulties of freezing blastocysts and involves only short periods of prolonged culture (18 hours). Finally, the transition towards automated classifiers offer continuous evaluation of embryonic developmental potential, thus providing the physicians with additional tools to decide when to transfer the embryos.

Methods

The SHIFRA database, embryo annotation and QA

We developed a PostgreSQL database with a front-end website that supports display, query and data annotation (Hebrew University IT). Maintenance of the SHIFRA database was outsourced (CHELEM LTD) and data curation was performed by a full-time trained embryologist. Anonymized time-lapse PDB video files and the corresponding metadata were imported from five hospitals across Israel. Data was imported under the approval of the Helsinki ethical committee of the Hebrew University and Hadassah medical center (IRB number HMO -006-20). Morphological and morphokientic annotations were performed by qualified and experienced embryologists in each IVF department and by a team of trained embryologists in the Buxboim lab adhering to established protocols. Quality assurance (QA) of morphokientic annotations was performed using randomly selected 277 embryos by a highly trained embryologist (Dr. IHV, Soroka medical center, Fig. S1).

Qualification of embryos for deep learning of BLAST and KID labeled embryos

The qualification criteria for including embryos in the BLAST and KID train/test and validation sets are summarized in the following table.

	BLAST		KID	
Set	**Train/Test**	**Validation**	**Train/Test**	**Validation**
ICSI	Yes	Yes	Yes	Yes
Fresh transfer	NA	NA	Yes	Yes
Discard PGD/PGS embryos	Yes	Yes	Yes	Yes
Two pronuclei	Yes	Yes	Yes	Yes
No dark / misaligned images	Yes	Yes	Yes	Yes
Full morphokientic annotation*	NA	Yes	NA	Yes

* tPNa to final morphokientic state.

Labeling of embryo blastulation

To identify BLASTn, we detect the embryos whose development is arrested based on the temporal distributions of the morphokinetic events and their intervals (Fig. S1a-b). Each embryo is represented by the latest morphokinetic state it reached inside the incubator, MS_n, total time

of incubation measured from fertilization, t_{inc} (Fig. 2b), and time that lapsed from the time of last morphokinetic event, $t_{inc} - t_n$. We represent each embryo by two Cartesian coordinates: $[MS_n, t_{inc}]$ and $[MS_n, t_{inc} - t_n]$ as plotted in figures 2a-i and 2a-ii, respectively. The light green regions in figure 2a-i correspond to the time windows of each morphokinetic event. To minimize the potential impact of statistical outliers, we bound these time windows between the 1st and the 99th percentiles of the morphokinetic distributions as measured based on thousands of embryos (Fig. 1c, Table S2). Therefore, embryos that are located in the light green regions exhibit normal development and can thus hold the potential to proceed in the case their incubation was extended. The dark green regions are upper-bounded by the 99th percentile of the next consecutive morphokinetic event, MS_{n+1}. Embryos that are located in these regions reached stage MS_n and did not proceed to the next morphokinetic event but are still within the permitted time window for proceeding to MS_{n+1}. On the other hand, embryos that are located in the red regions have passed their time window and are thus arrested in morphokinetic stage MS_n. For example, an embryo that was incubated for 96 hours and reached 4-cells state, is arrested and is defined as 4-cells positive 5-cells negative ($4C_p5C_n$).

A similar statistical analysis is performed to study the temporal distributions of the morphokinetic intervals (Fig. 2a-ii). The light green regions, corresponding to the time interval windows, are bound by the 1st and 99th percentiles of the time interval distributions (Fig. 1d, Table S3). Embryos that are located in the red regions passed their interval time window and are thus arrested in morphokientic stage MS_n. For example, an embryo that had reached the 4-cells stage and that 36 hours have lapsed since this cleavage event without completing the next cleavage, is defined $4C_p5C_n$ and is arrested. Adhering to a constringent approach, embryos that were found to be developmentally arrested based on the timing of their morphokinetic events or their morphokinetic profiles are classified BLASTn. The remaining embryos are classified BLAST-unknown (BLASTu, Fig. 2c-i,ii).

Preprocessing of embryo video files

Automated screening empty well images

We determine if an image contains an embryo or the culture well is empty by applying the following algorithm. Horizontal and vertical 3x3 Prewitt operators are applied on the input image (Fig. S3a-i). In each pixel, the L^2-norm of the absolute values of both channels is calculated to generate a gradient map which is then normalized by the median gradient value. To find the boundaries of the circular culture well, we treat each side of the image separately (left, top, right, bottom) as illustrated below for the top side of the image. The central fifty columns (columns 226 to 275) are scanned from top to bottom and the first pixel with value greater than 0.4 is marked in each column. Next, the average y-axis coordinate of all fitty marked pixels is calculated. By applying these steps also to the bottom, left and right sides we define the rectangular in which the culture well is bounded. The boundaries of the well are obtained by fitting a circle inside. Then, all the pixels that are located outside the well boundaries are set to zero (I_1, Fig. S3a-ii). Low gradient levels are removed by setting low-intensity pixels $I_1 < 0.2$ to zero followed by removing 'salt and pepper' noise using 10x10 convolution mask (I_2, Fig. S3a-iii). The remaining objects in I_2 are typically larger than the size of the convolution mask. Under the assumption that the embryo is the largest object with the sharpest edges, we set to zero all pixels except of the pixels that belong to the highest-energy cluster, namely the object with the maximal integrated intensity in I_2 (I_3, Fig. S3a-iv). An image is labelled empty if the sum of I_3 pixels is lower than 8000.

Automated selection of the sharpest focal plane

The selection of the sharpest focal plane out of all seven z-stacks is performed as described in the section above ('Automated screening empty well images'). The I_3 images are calculated for all focal planes and the frame with highest score is selected to be the sharpest focal plane.

Automated cropping and down-sampling of the ROI's of embryos

Automated segmentation of the embryo's ROI is illustrated in figure S3b. First, a self-similarity descriptor (SSIM) is applied on the input image (Fig. S3b-i) to generate a gradient map, I_1. Briefly, gradients are calculated at each pixel along eight equally-rotated directions (45° apart) at a distance of three pixels away. For each angle, noise is reduced by averaging the gradients over 3x3 regions centered at distal pixels. Next, the L^2-norm of the all eight directional gradients is calculated and the value of each pixel in I_1 is obtained by normalizing by the median SSIM value of the entire image. The SSIM depicts and highlights the edges of the well and of the embryo (Fig. S3b-ii). Next, we perform a convolution between I_1 and a 256x256 mask of ones. Since I_1 is a 500x500 matrix, the product of this convolution, I_2, is 245x245 (Fig. S3b-iii). The location of the region of interest (ROI) in which the embryo is contained is obtained by the argument of maxima (ArgMax) of I_2 (Fig. S3b-iv). Finally, the cropped ROI was down-sampled two-fold biaxially (128x128 pixels).

Grouping of frame packets

Each packet P_n^m was defined by a binary label, y_m, according to the label of embryo m (BLASTp/BLASTn or KIDp/KIDn) and by a time index n of the middle frame (Fig. 3a).

CNN training

Network architecture

Network design consisted of three convolution-max pooling layers, followed by two residual-max pooling convolution blocks, five residual convolution blocks, and two fully-connected layers (Fig. 3b). The final fully-connected layer generates six values that correspond to six 12-hours' time windows ranging between 2 to 5 days (packets earlier than 48 hours were included in the first time window). Final packet score is chosen out of these six values based on the time index.

Training parameters

The input of the network was time-lapse video files where each training batch consisted of K embryos. For simplicity, here we describe the training process of $K = 1$. To reach network convergence, each embryo was represented by randomly selected k packets within a 20 hours' time window such that pairs of packets were separated by no longer than 8 hours. Training set embryos were represented using randomly selected packets of all focal planes whereas test and validation set embryos were represented using packets only of the sharpest focal plane (see above). Each embryo was thus represented by an object of size 128x128x5xk. In practice, we used $K = 8$ batch size where single embryos were represented by $k = 4$ packets. Compared with 500x500x200 pixels for a typical video file, the size of the input object was decreased by 150-fold as part of preprocessing.

Modified soft hinge loss

The weighted soft hinge loss of embryo m is calculated based on all of its k *packets*:

$$l^m = \sum_{n=1}^{k} w_n^m Log\left(1 + e^{-y_m \cdot score_n^m}\right),$$

where $score_n^m$ is the packet score and w_n^m is the adaptive weight:

$$w_n^m = \frac{e^{\gamma * score_n^m}}{\sum_{n=1}^{k} e^{\gamma * score_n^m}}$$

Note that the sum of weights w_n^m for embryo m is 1. γ is the softness parameter. At the limit $\gamma \to 0$, the weight becomes $\frac{1}{k}$ independent of the scores of the packets. At the opposite limit of large γ, w_n^m approaches 1 only for the packet of maximal score. The problem of approaching this limit is that it will be increasingly difficult for the network to converge. The batch loss L is:

$$L = \frac{1}{K} \sum_{m=1}^{K} l^m$$

The CNN weights are thus optimized to minimize L. Performances were optimized by setting $\gamma = 5$. For a negative embryo ($y_i = -1$), even if a single packet will obtain a positive score, its weight w_n^m will be highest and the loss of the embryo l^m will be large. As a result, convergence will be

approached only if all packets of a negative embryo will obtain negative scores. On the other hand, one packet with a high positive score is sufficient for obtaining a small loss for a positive embryo ($y_i = +1$). For these embryos, the packets with low scores will have small weights and the packet with the highest score will have the highest weight and a small loss will be obtained.

Integrating packet scores into embryo scores

Embryos were scored based only on the packets that belong to the twelve hours that preceded the time of classification. With 18 to 20 min time-lapse intervals, each embryo was thus scored based on 36 to 40 packets. We introduced high temporal resolution during the process of integrating packets' scores into an embryo score by dividing the packets from all embryos into separate cohorts based on a grid of two-hours' time windows. Unlike BLAST-labeled packets, which obtained a single value by the network, KID-classification involved six subnetworks that generated six implantation scores as described in the main text. Hence, we included an additional step for integrating KID-labeled packet scores as follows. Within each time-window, a soft SVM with a linear kernel was trained on packets of all training set embryos (6 to 8 packets per embryo). To integrate packets' scores into an embryo score, threshold values were introduced and optimized for each two-hours' time window by maximizing the AUC as computed over test set embryos. Packets that were scored lower than threshold were discarded. Embryos were thus scored by summing up only high-score packets after subtracting the corresponding threshold values. In this manner, BLAST-prediction and KID-prediction of embryos of high developmental competence were performed.

Calculation goodness of classification

Let ρ_n^m be the score of packet P_n^m of embryo m at time n. Within a time window W_t, the average packet score of embryo m is:

$$\hat{\rho}_m(t) = \frac{1}{|\{n \in W_t\}|} \sum_{n \in W_t} \rho_n^m$$

Denoting the positively (negatively) labeled embryos by m^+ (m^-), the goodness of classification of time window W_t is defined by:

$$\alpha(t) = \frac{\frac{1}{M^+(t)} \sum_{m^+=1}^{M^+(t)} \quad \widehat{\rho}_{m^+}(t) - \frac{1}{M^-(t)} \sum_{m^-=1}^{M^-(t)} \quad \widehat{\rho}_{m^-}(t)}{\frac{1}{M(t)} \sum_{m=1}^{M(t)} \quad \left|\widehat{\rho}_m(t)\right|}$$

Here, $M(t)$ is the number of embryos with packets in time window W_t. $M^+(t)$ and $M^-(t)$ correspond to positively and negatively labeled embryos, respectively. $\alpha(t)$ obtains high values in the case that positively-labeled packets are scored high and negatively-labeled packets are scored low.

References

1 Prevention, C. f. D. C. a. *ART Success Rates*, <https://www.cdc.gov/art/artdata/index.html> (May 16, 2018).

2 Guerif, F. *et al.* Limited value of morphological assessment at days 1 and 2 to predict blastocyst development potential: a prospective study based on 4042 embryos. *Hum Reprod* **22**, 1973-1981, doi:10.1093/humrep/dem100 (2007).

3 Alpha Scientists in Reproductive, M. & Embryology, E. S. I. G. o. The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Hum Reprod* **26**, 1270-1283, doi:10.1093/humrep/der037 (2011).

4 American College of, O., Gynecologists Committee on Practice, B.-O., Society for Maternal-Fetal, M. & Committee, A. J. E. ACOG Practice Bulletin #56: Multiple gestation: complicated twin, triplet, and high-order multifetal pregnancy. *Obstet Gynecol* **104**, 869-883 (2004).

5 Hlinka, D. *et al.* Time-lapse cleavage rating predicts human embryo viability. *Physiol Res* **61**, 513-525 (2012).

6 Nakahara, T. *et al.* Evaluation of the safety of time-lapse observations for human embryos. *J Assist Reprod Genet* **27**, 93-96, doi:10.1007/s10815-010-9385-8 (2010).

7 Meseguer, M. *et al.* The use of morphokinetics as a predictor of embryo implantation. *Hum Reprod* **26**, 2658-2671, doi:10.1093/humrep/der256 (2011).

8 Ciray, H. N. *et al.* Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. *Hum Reprod* **29**, 2650-2660, doi:10.1093/humrep/deu278 (2014).

9 Petersen BM, Boel M, Montag M, Gardner DK. Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3. Hum Reprod. 2016 Oct;31(10):2231-44. doi: 10.1093/humrep/dew188. Epub 2016 Sep 8. PMID: 27609980; PMCID: PMC5027927.

10 Shalev-Shwartz, S. S., Shaked; Shashua, Amnon. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving.

11 Suk, H. I., Lee, S. W., Shen, D. G. & Initi, A. s. D. N. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage* **101**, 569-582 (2014).

12 Greenspan, Y. B. I. D. L. W. S. L. E. K. H. in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* (2015).

13 Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun Acm* **60**, 84-90, doi:10.1145/3065386 (2017).

14 Liu, W. W., Yandong; Yu, Zhiding; Li, Ming; Raj, Bhiksha; Song, Le. SphereFace: Deep Hypersphere Embedding for Face Recognition.

15 Schroff, F. K., Dmitry; Philbin, James. FaceNet: A Unified Embedding for Face Recognition and Clustering.

16 Howard, A. G. Z., Menglong; Chen, Bo; Kalenichenko, Dmitry; Wang, Weijun; Weyand, Tobias; Andreetto, Marco; Adam, Hartwig. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

17 Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vision* **115**, 211-252, doi:10.1007/s11263-015-0816-y (2015).

18 Avendi, M. R., Kheradvar, A. & Jafarkhani, H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med Image Anal* **30**, 108-119, doi:10.1016/j.media.2016.01.005 (2016).

19 Cheng, J. Z. *et al.* Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Sci Rep-Uk* **6**, doi:ARTN 24454 10.1038/srep24454 (2016).

20 Ronneberger, O. F., Philipp; Brox, Thomas. U-Net: Convolutional Networks for Biomedical Image Segmentation.

21 Swain, J.E. et al. (2016). "Optimizing the culture environment and embryo manipulation to help maintain embryo developmental potential." *Fertility and Sterility* 105(3):571 – 587.

22 Chen F, De Neubourg D, Debrock S, Peeraer K, D'Hooghe T, Spiessens C. Selecting the embryo with the highest implantation potential using a data mining based prediction model. Reproductive Biology and Endocrinology. 2016; 14: 10. doi:10.1186/s12958-016-0145-1.

23 Lemmen, J. G., Agerholm, I. & Ziebe, S. Kinetic markers of human embryo quality using time-lapse recordings of IVF/ICSI-fertilized oocytes. *Reprod Biomed Online* **17**, 385-391, doi:Doi 10.1016/S1472-6483(10)60222-2 (2008).

24 Zhan, Q., Ye, Z., Clarke, R., Rosenwaks, Z. & Zaninovic, N. Direct Unequal Cleavages: Embryo Developmental Competence, Genetic Constitution and Clinical Outcome. *PLoS One* **11**, e0166398, doi:10.1371/journal.pone.0166398 (2016).

25 Wang, X. *et al.* Conception, early pregnancy loss, and time to clinical pregnancy: a population-based prospective study. *Fertil Steril* **79**, 577-584 (2003).

26 Wong, C. C. *et al.* Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nat Biotechnol* **28**, 1115-U1199, doi:10.1038/nbt.1686 (2010).

Supplementary figures

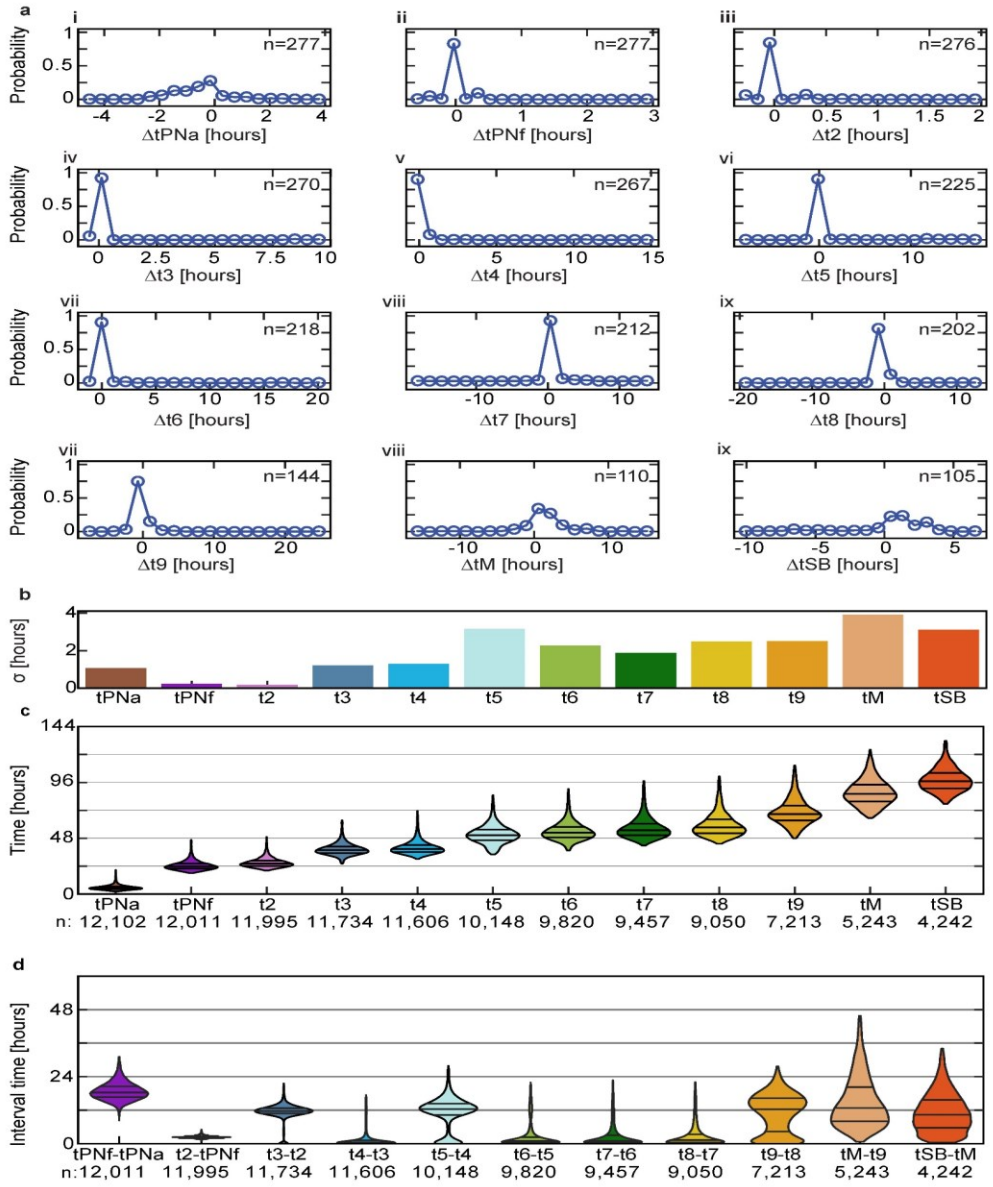


Figure S1 | High-resolution morphokientic distributions obtained based on thousands of embryos and validated by quality assurance. (a) Morphokinetic distributions show partial temporal separation between first (tPNf-t2), second (t3-t4) and third (t5 to t8) cleavage cycles. (b) Compared with the dynamics within each cell cycle, distributions of consecutive intervals show slow transition between cell cycles. (c) Quality assurance (QA) of morphokinetic annotation was performed by an expert and experienced embryologist based on established protocols. Histograms of the differences between SHIFRA annotations and QA measured for a representative subset of embryos (277 embryos). (d) Standard deviations of the morphokinetic differences are shown. Note that time-lapse intervals are 20 min.

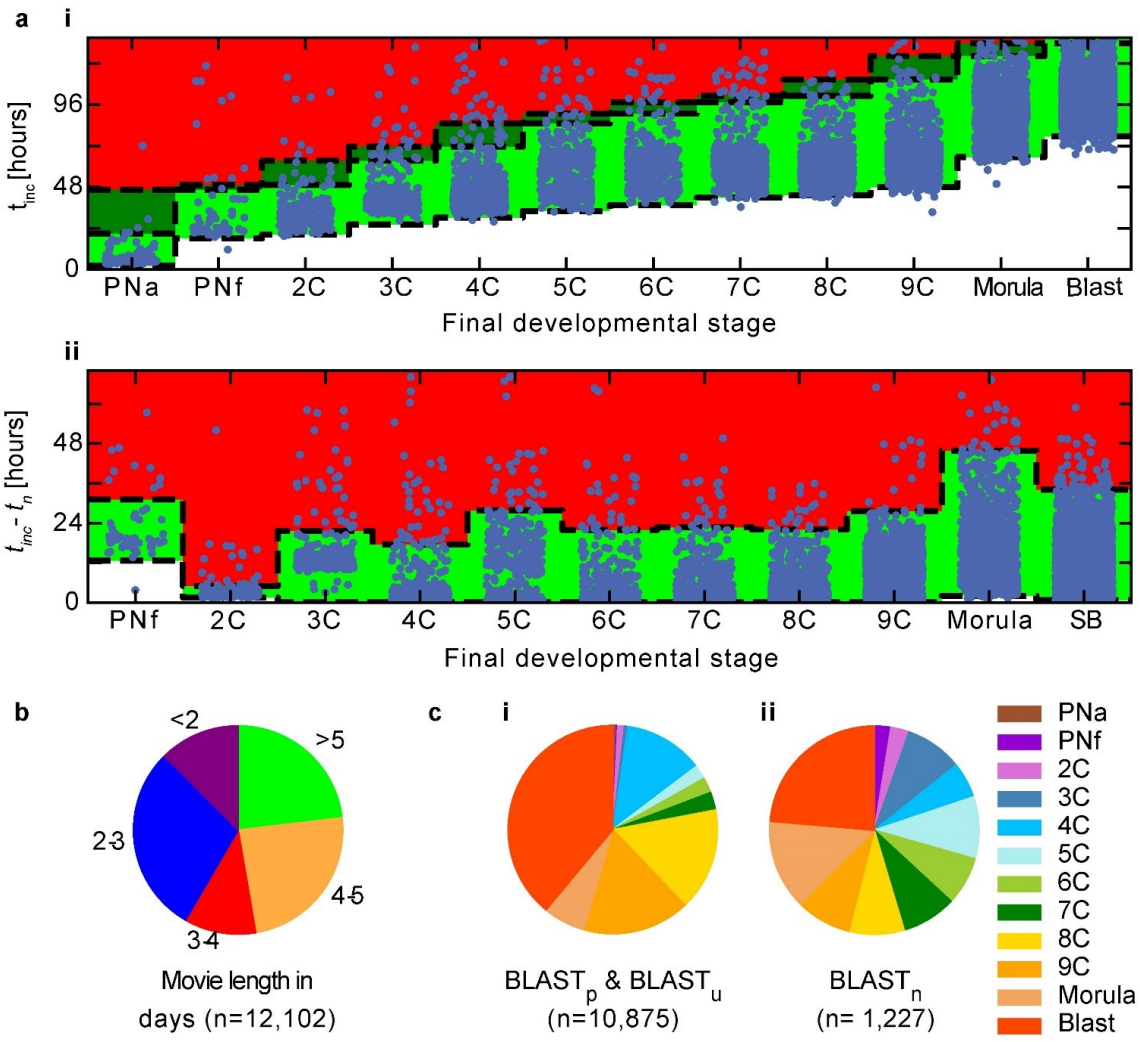


Figure S2 | Labeling of blastocyst-negative (BLAST$_n$) embryos. (a-i) Phase diagram separating between the embryos (blue dots) that have the capacity to proceed further to the next stage in their preimplantation development (green regions) and the embryos that are arrested at the specified stage (red region) based on their incubation time (t_{inc}) and their final developmental stage reached inside the incubator. The morphokinetic time windows are marked light green, bounded by the measured 1st percentile (bottom dashed line) and 99th percentile (middle dashed line) of morphokinetic distributions (Fig. 1c). Top dashed line marks the 99th percentile of the next morphokientic event. (ii) Phase diagram based on the morphokinetic intervals that lapsed from time of most advanced morphokinetic stage that the embryos reached (t_n). Light green regions mark the time windows of morphokinetic intervals, bounded by the 1st percentile (bottom dashed line) and the 99th percentile (top dashed line) calculated based on morphokinetic interval distributions (Fig. 1d). Embryos that appear in the red region of either i or ii are labeled BLAST$_n$. Distributions of (b) embryo video length and (c) most advanced morphokinetic state of BLAST-labeled embryos (color coded, right bar). The SHIFRA database consist of **4522** BLAST$_p$ and **1489** BLAST$_n$ embryos.

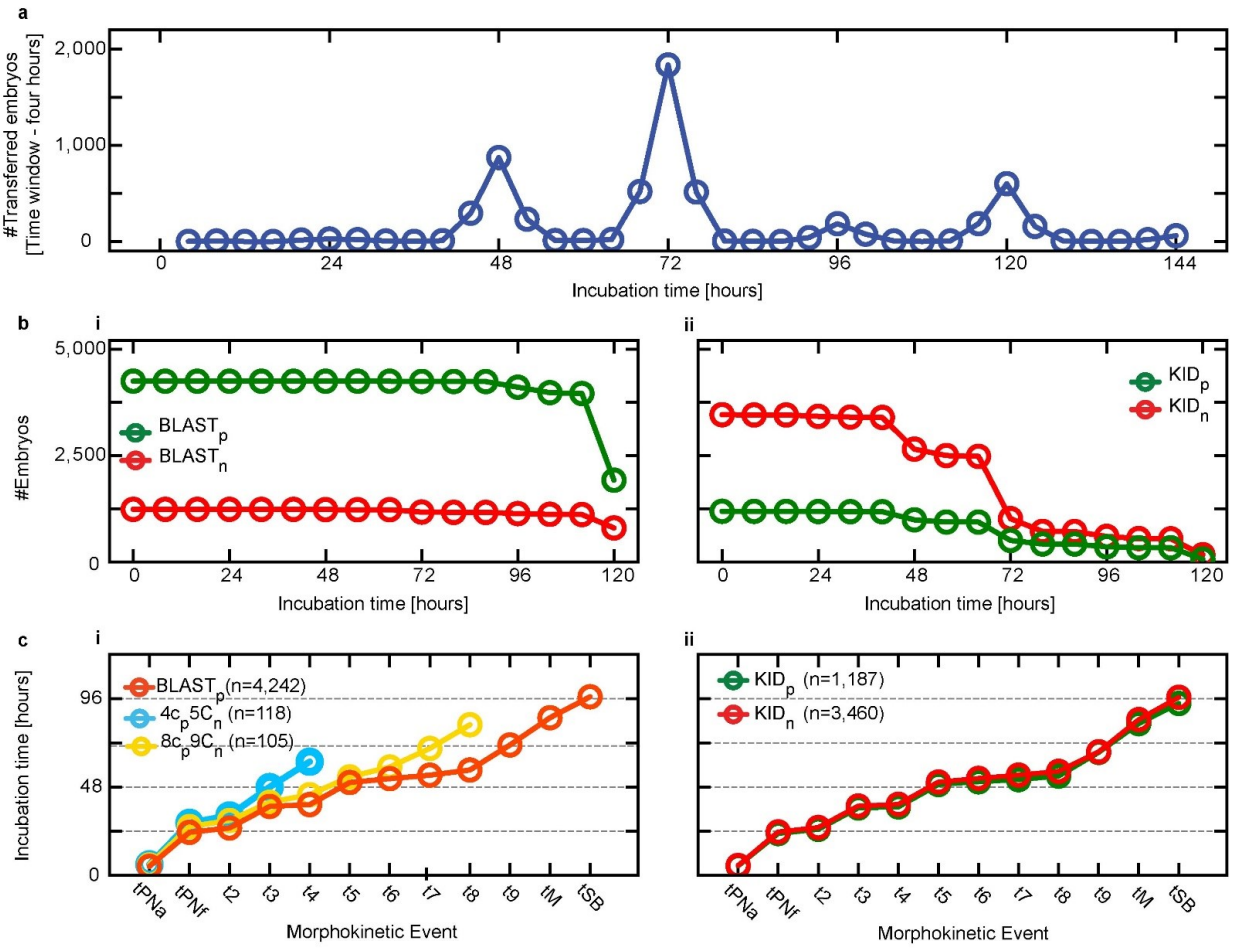


Figure S3 | Embryo transfer and implantation statistics. (a) Embryos are transferred to the uterus on days 2, 3, 4, and 5. (b) Number of (i) BLAST- and (ii) KID-labeled embryos decrease with incubation time. The number of KID_p and KID_n embryos decreases fast in correlation with time of embryo transfer (shown in a). (c) The median morphokinetic trajectories of (i) embryos that were arrested at 4-cells stage (4 cell-positive 5 cell-negative; 4Cp5Cn) and embryos that were arrested in 8-cells stage (8Cp9Cn) are compared with $BLAST_p$ embryos, and (ii) KID_p versus KID_n embryos demonstrating complete overlap.

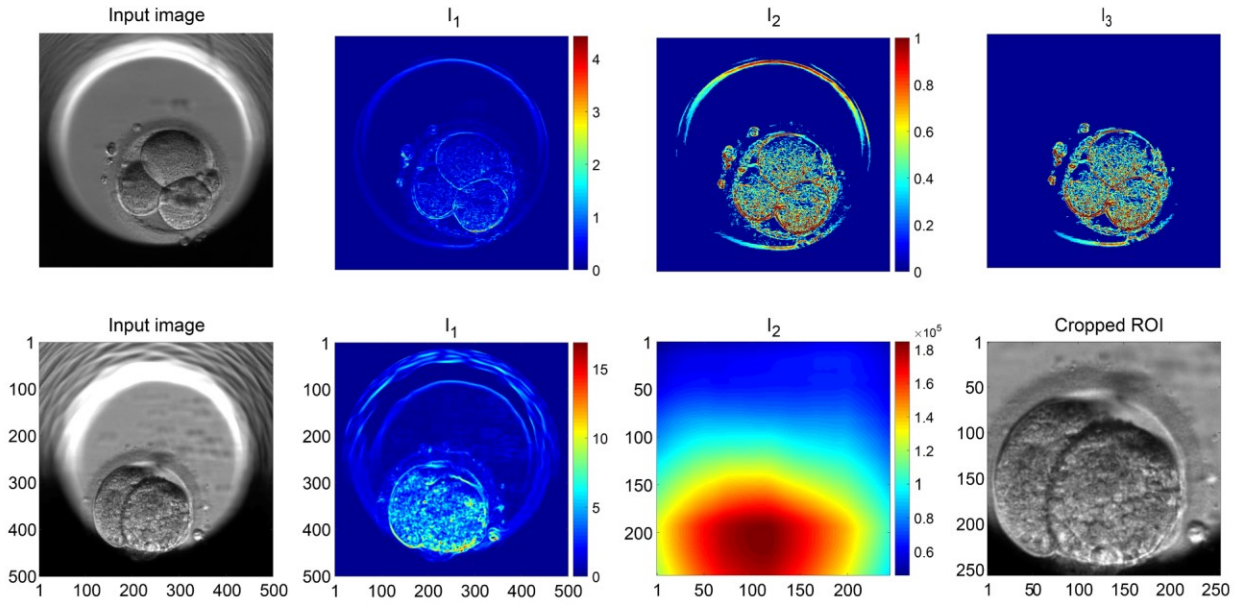


Fig S4 | Screening empty well images, defining sharpest focal plane frame and cropping embryo ROI: (a) Empty well images are detected as illustrated here for a 3-cells input image. A gradient map is calculated, the boundaries of the well are segmented and external pixels are set to zero (I_1). After setting low-intensity pixels to zero and removing noise via smoothing (I_2), the largest object which is the embryo is segmented (I_3). Images are set empty if the integrated intensity of I_3 is lower than a threshold intensity 8000. The sharpest focal plane frame is defined by the maximal integrated intensity. (b) Embryo segmentation is demonstrated for a 2-cells embryo input image. An SSIM descriptor generates a gradient map (I_1). Next, a 256x256 pixels convolution mask is applied, which highlights high-gradient regions (I_2). Based on the texture of the cytoplasm and the surrounding zona pellucida, the coordinate of the ROI mask which contains the embryo (cropped ROI) is obtained at the maximal intensity pixel of I_2.

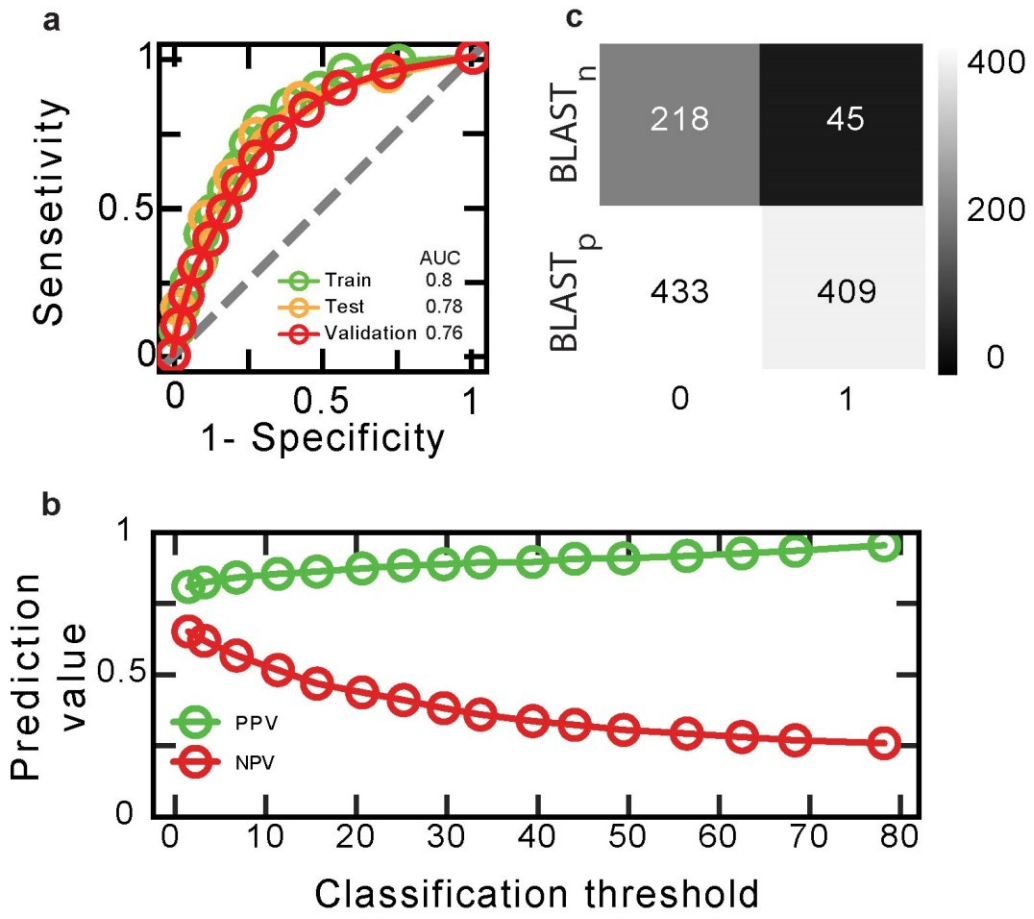


Figure S5 | Characteristics of the SHIFRA$_B$ classifier. (a) Overlapping ROC curves of the train, test and validation sets indicates lack of overfitting. (b) Positive and negative predictive values (PPV and NPV) are plotted as a function of the classifier's threshold at 72 hours. (c) Confusion matrix at classifier threshold of 40.7 accounts for PPV=0.9 and NPV=0.33, thus demonstrating clinically-relevant classification performances by SHIFRA$_B$ at 72 hours.

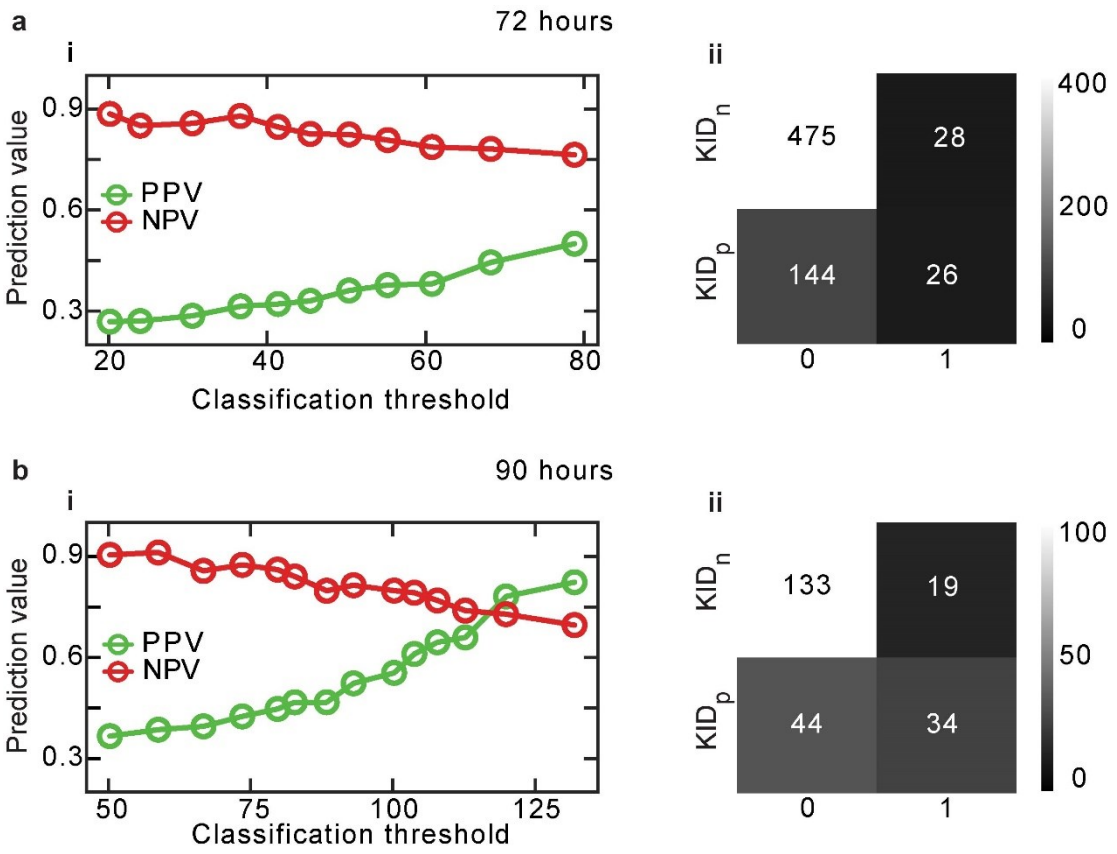


Figure S6| Characteristics of the SHIFRA$_K$ classifier. PPV and NPV are plotted as a function of threshold at (a) 72 and (b) 90 hours. The confusion matrices illustrate SHIFRA$_K$ performances at the following threshold values: (a-ii) 77. (b-ii) 110.

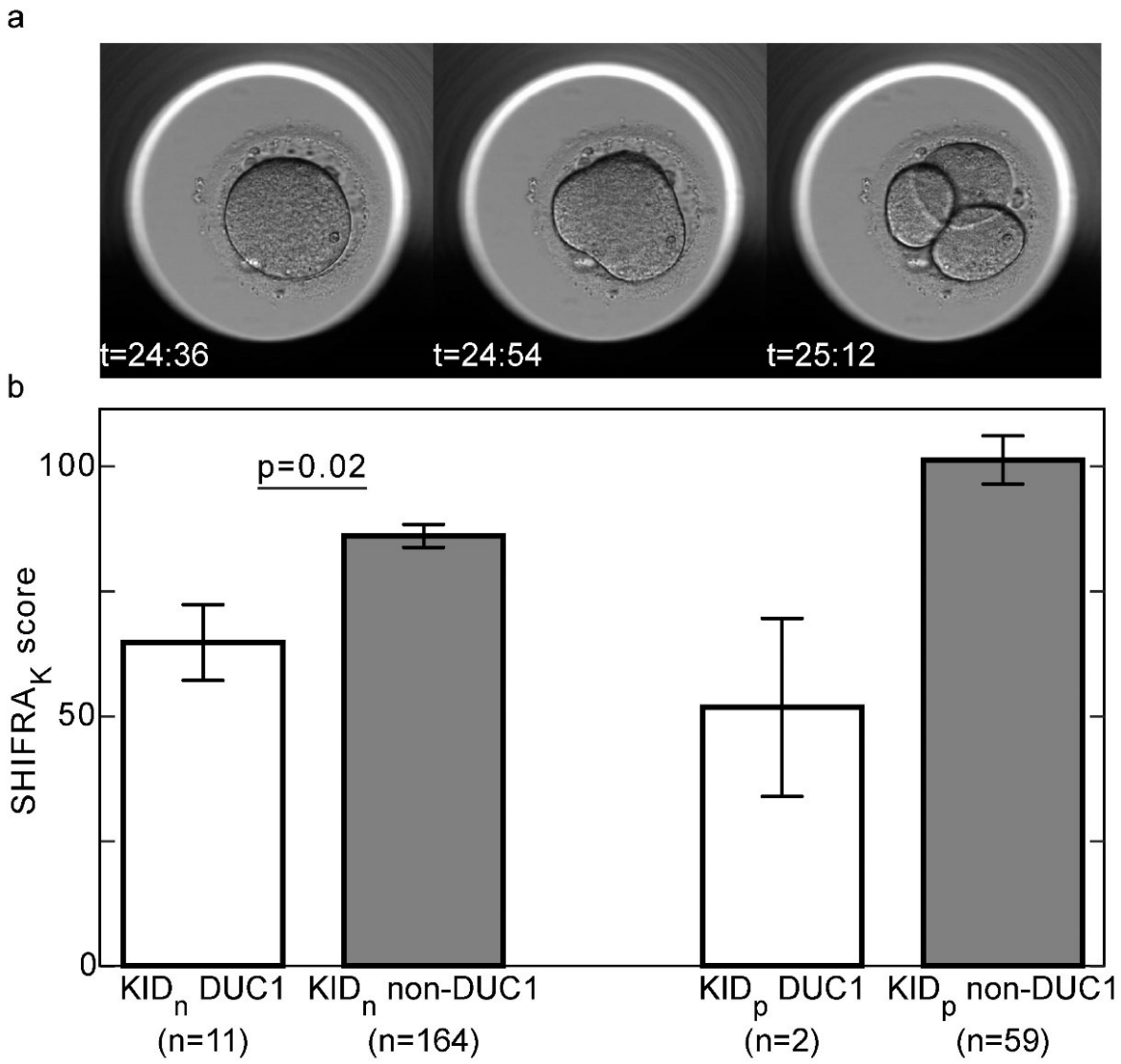


Figure S7 | DUC1 embryos are deselected by SHIFRA$_K$. (a) First direct unequal cleavage (DUC1) defines the direct cleavage of the first cell into three cells. (b) DUC1 embryos are scored low relative to non-DUC1 embryos by SHIFRA$_K$. This is true for KID$_n$ and KID$_p$ embryos. Due to their small number, analysis included all DUC1 embryos in the SHIFRA database compared with validation-set non-DUC1 embryos. P-value could not be evaluated for KID$_p$ embryos due to the small number of DUC1 embryos.

Evaluating implantation prediction models: A comparative analysis of machine learning algorithms using morphology versus morphokinetics

Itay Erlich[1,2], Iris Har-vardi[2,3], Michelle Tran[2], Gilad Karavani[4], James A. Grifo[5], Fang Wang[5], Assaf Ben-Meir[2,4]

[1] The Alexender Grass Center for Bioengineering, School of computer Science and Engineering, Hebrew University of Jerusalem, Israel

[2]Fairtilty Ltd., Tel Aviv, Israel

[3]Fertility and IVF unit, Department of Obstetrics and Gynecology, Soroka University Medical Center and the Faculty of Health Sciences Ben-Gurion University of the Negev, Beer-Sheva, Israel

[4]IVF unit, Department of Obstetrics and Gynecology, Hadassah Medical Organization and Faculty of Medicine, Hebrew University of Jerusalem, Israel

[5]New York University Langone Prelude Fertility Center, New York, New York

[*] Submitted to *human reproduction*

Abstract

Objective: To validate implantation prediction models using artificial intelligence algorithms (AI) to compare results from morphology assessment data, automatic morphokinetics events evaluation or combination of both data in a retrospective multi-center study.

Design: Retrospective cohort study

Setting: Three University-affiliated IVF centers

Patients: This study included 36561 embryos' development videos obtained from time-lapse incubators between 2014 – 2019.

Intervention: None

Main Outcome Measure (s): The automatic morphokinetic evaluation tool was trained on 36561 annotated embryo development videos. Morphokinetic annotation and morphology evaluation of 6938 embryos with known implantation data (KID) were used to train and test KID positive (KID+), an AI algorithm for implantation prediction. The training set consisted of 6363 embryos (1078 KID+ and 5285 KID-negative (KID-). The blind test set consisted of 575 embryos (171 KID+ and 404 KID_). The embryos were scored for implantation potential by KID+ based on automatic evaluation of morphokinetic and morphology data. We compared our combined morphokinetic and morphology model to models that only take in morphokinetic events or the last frame of embryo development.

Results: In this study we demonstrated a machine learning algorithm that automatically annotates embryo morphokinetic events (r^2=0.95). Incorporation of these estimated

annotations into a CNN algorithm revealed a robust implantation prediction tool with an area under curve (AUC) continuously improving every hour from 30 to 116 hours post insemination, reaching a maximal AUC of 0.65 at 116 hours. However, analysis of the last frame of the embryo development after morula stage showed better prediction with AUC of 0.68 at 116 hours. Combining both algorithms revealed that morphokinetics annotation added value until the start of blastulation (around 90 hours following insemination). After the start of blastulation, the combined tool was similar to morphology alone with AUC of 0.68 at 116 hours.

Conclusion: Assessment of blastocyst morphology is sufficient for most of its competence evaluation. However, prediction at cleavage stage is better when incorporating morphokinetic data into morphology evaluation.

Keywords: Artificial intelligence, embryo selection, time-lapse image, morphokinetics annotation, blastocyst morphology

Introduction

The goal of fertility treatment is to deliver a healthy baby. Multiple embryo transfer was common practice to increase the pregnancy rate but subsequently also increased multiple pregnancy prevalence and the complications associated with multiple pregnancies. Multiple pregnancy increases the likelihood for preterm deliveries and consequently exposes the newborn to prematurity complications, such as intraventricular hemorrhage, respiratory distress syndrome and necrotizing enterocolitis [1]. Hence, the way to reach singleton pregnancy is to adopt the policy of elective single embryo transfer [2]. However, the poor predictive power of existing tools to select the most viable embryo encourages us to transfer more than one embryo to achieve a reasonable pregnancy rate. Therefore, a robust non-invasive predictive tool for embryo selection is mandatory.

Better methods for assessing embryo quality have been a long-term goal in assisted reproductive technology (ART). Morphology grading, namely the Gardner score, is a non-invasive method and routinely used in clinical practice for blastocyst selection [3]. The transfer of the top scoring blastocyst has been shown to achieve the highest implantation rates [4, 5]. However, morphology grading has been known to have significant intra- and inter-observer variability [6].

In recent years, time-lapse incubators (TLI), which image the embryo every 15 – 20 minutes, entered in-vitro fertilization labs and clinics [7]. TLI decrease the work burden on embryologists by eliminating the need to remove the slides from the incubator to observe embryo development and grade. Embryologists can review the generated movie and navigate back to specific developmental events to check for morphokinetic or morphological abnormalities, such

as direct unequal cleavage (DUC) or reverse cleavage, phenomena observed in low implantation potential embryos [8]. Moreover, TLI have enabled morphological and morphokinetic event data (timing of every milestone in embryo development) to be captured, thus presenting the opportunity to develop an algorithm to predict embryo implantation potential [9-11].

Several algorithms that analyze embryo kinetics and morphology have been proposed to predict blastocyst formation [12], genetic chromosomal disorders [13, 14] and implantation potential [15],. Some analyze morphology and morphokinetic parameters from time lapse images [13, 16, 17] but require manual annotation, thus are associated with intrinsic user variability [6, 16].

Deep Convolutional Neural Networks (CNNs) are deep learning algorithms that can take in an input image, assign importance to various features in the image, and differentiate one from the other. In particular, CNNs can exploit both spatial and temporal information, possibly at the same time, using 3D convolutional layers or recurrent networks [17]. CNNs have led to state-of-the-art results in a variety of problems, such as image classification and segmentation, object detection, video processing, speech recognition and natural language processing [18, 19]. Inspired by how the human brain works, deep CNN uses multiple feature extraction stages that can automatically learn abstracted representations from the data. The growing availability of data and advances in computing power have accelerated research in CNNs.

As mentioned above, embryo prediction tools are divided into morphokinetic-based algorithm [20] or morphological-based single image recognition [21, 22]. In this work, we aim to compare algorithms for implantation prediction based on morphology alone, morphokinetics,

or the combination of both. We investigated how the morphology information extracted from the last frame of a developmental video compares to the information extracted from the course of the entire development, which includes morphokinetic events, can be used to predict implantation.

Method

Data collection

Clinical and IVF laboratory data was collection from three units including Hadassah Hebrew University Medical Center Ein-Kerem and Mt Scopus campuses, , and New York University Langone Prelude Fertility Center. Each unit uses their own controlled ovarian hyperstimulation protocols, although most of the cycles were antagonist protocols. After fertilization, the embryos entered into an Embryoscope^{TM} (Vitrolife, Copenhagen, Denmark), a TLI that captures a seven-layer z-stack images 15 µm apart at each time point every 15 to 20 minutes. The patient pool was heterogeneous and included patients with a variety of infertility indications. Patients with preimplantation genetic testing (PGT) cases were excluded due to the fact that PGT requires the embryo to be removed from the incubator for the biopsy, which may interrupts the subsequent morphology and morphokinetics of embryo development. The study was approved by the Investigation Review Board of Hadassah Hebrew University Medical Center (IRB number HMO -006-20).

Embryo classification and morphokinetic annotation

The outcome of transferred embryo was divided into 3 groups: 1. KID-positive: the numbers of gestational sacs equal the number of embryos transferred. 2. KID-negative: none of the embryo transferred reached a gestational sac stage. 3. KID unknown: The number of gestational sac (at least one) is lower than the number of transferred embryos. Embryonic manual annotation was carried out by three expert embryologists, and the definition of each developmental event was based on common nomenclature [23].

Preprocessing

The input data of each embryo consists of a series of z-stack images. The n^{th} image of embryo *i* is denoted by $\{I_n^i\}$. Embryo *i* consists of N_i images and each image I_n^i is a 3-dimensional file of size 500x500x7 pixels (corresponding to 7 focal planes).

The spheroid culture well is centered in each image (500x500 pixels), but the embryo itself can be found anywhere inside the well (430x430 pixels). In order to minimize systematic variation between images, we cropped the region of the embryo inside the well. This is performed using a CNN model that recognizes and segments the embryo inside an image. The CNN architecture is outlined in Fig 1A. Finally, the segmented embryo is cropped from the original image and resized to an 128x128 pixels image.

Deep learning for automatic annotation

The Morphokinetic State (MS) is defined as one of the following 15 states namely zygote, time of pronuclei appearance (tPNa) and fading (tPNf), time of 2 distinct cells (t2) to 9 cells and above (t9+), time to morula (tM), time to Start Blastulation (tSB), time to Blastocyst (tB) and time to Expanded Blastocyst (tEB).

The label of an image with morphokinetic state *k* is denoted by $y_k, k = [1 \ldots 15]$.

A CNN model is first trained to classify each image by its respective morphokinetic state. Namely, the training data consists of pairs of $\{I_n^i, y_n^i\}$ where I_n^i denotes an embryo cropped image at size 128x128 pixels of embryo *i* taken at time *n,* and y_n^idenotes its morphokinetic label. The model accepts the image $I_n^i$and outputs the estimated morphokinetic state $\widehat{y_n^i}$. This learning is performed by minimizing the multi-class cross entropy loss function. The CNN is implemented by the commonly used ResNet50 architecture which outputs a vector with 15 dimensions (s^n) populated with the probability that the input image exists in each state (Fig. 1B). Morphokinetic annotations are defined by the start of each of the morphokinetic states. To estimate the annotations given the per-frame soft probabilities, we used dynamic programming. At each time n and for each state k, the best paths at time n of the states { k -1, k } are compared. The best path of the k-th state at time n +1 is the path with the highest score at time n, concatenated with the state k at time +1. The new path score is comprised of the best path from the { k, k -1} states score and the prediction score at time n +1 of the k-th score (s_k^{n+1}).

Implantation prediction based on morphokinetic annotation

To predict implantation outcome based on the estimated morphokinetic annotations, we use a simple Neural Network (NN) that accepts the estimated morphokinetic vector. The NN consists of six consecutive dense layers, followed by a scalar output layer, which represents the

implantation probability. The training data consists of a pair of $\{T_n^i, y^i\}$, where T denotes a 15 dimensional vector with the estimated annotations from all images till time *n* (where inf denotes annotations that didn't appear up to time *n*) and $y^i = \{1, -1\}$ denotes the label of embryo *i* corresponding to KIDp/KIDn. The architecture of the NN is given in Fig 1C.

Implantation prediction based on a single image

Here, implantation prediction is obtained from single images rather than the morphokinetic vector. We use a CNN that accepts a single image to predict implantation score. The training data consists of a pair of $\{I_n^i, y^i\}$, where I_n^i is the cropped image of embryo *i* taken at time n, and y^i remains as defined above. The CNN is implemented by ResNet50 architecture as illustrated in Fig 1D.

Implantation prediction based on annotations and single image

We combine the two basic morphokinetic and morphological models into a single network that is trained simultaneously. The network accepts both images and morphokinetic annotations and outputs the predicted implantation probability. Training data consists of $\{I_n^i, T_n^i, y^i\}$, all of which are as defined above. The architecture of this network is simply obtained by the composition of both basic architectures presented in Fig 1E. We did not incorporate patients' age into the algorithm in order to assess the contribution of morphology or morphokinetics on implantation prediction without the bias of age.

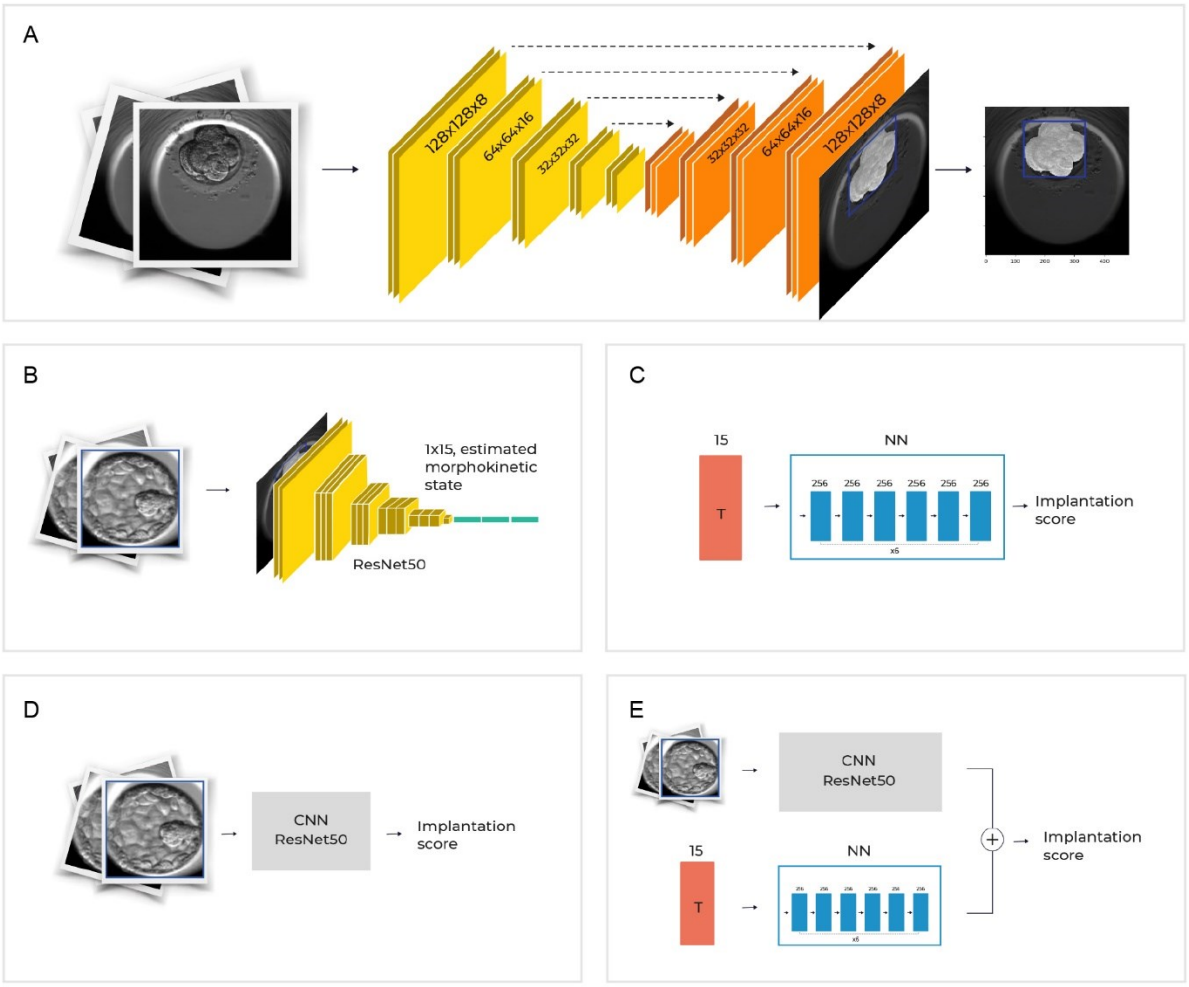


Fig 1: (A) Embryo segmentation CNN: Each image (500x500 pixels) is resized to 128x128 pixels and fed to a network with U-Net architecture. The output is a binary pixel-wise mask of the embryo which is then bounded by a rectangle. Finally, the cropped image is obtained by cropping the embryo part from the original image and resizing to 128x128 pixels. (B) Embryonic morphokinetic classification - A Cropped image is fed to ResNet50 CNN. The output is a 15 vector with the probabilities to be in each state. Finally, the estimated morphokinetic state is obtained by argmax over the probabilities output vector. (C) Implantation binary classification based on morphokinetic annotations - An input estimated annotations are processed by 6 dense layers to yield implantation probability. (D) Implantation binary classification based on single images. An input image is processed via a network with ResNet50 architecture to yield implantation probability. (E) Implantation binary classification based on integrating both morphokinetic annotations and morphology models.

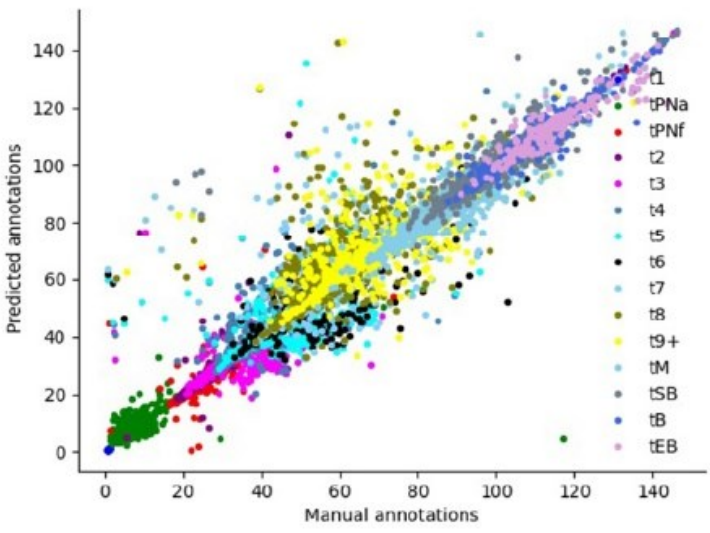
Results

Data collection and manual annotation

The embryo development videos were collected from July 2014 until December 2019 from three IVF clinics. Three senior embryologists supervised the annotation of 36561 embryos. The average age of women included in this database was 33.1±6.0 years (table 1). Data on implantation results was known for 6938 embryos with 1249 KID-positive embryos and 5689 KID-negative embryos.

Automatic annotation

All 36561 of the annotated embryos were used in this study with 34132 embryos for training and 2429 embryos for testing. Figure 2 illustrates the predicted annotation compared to the manual annotation of 15 morphokinetic events from the zygote stage to the time of expanded blastocyst. The overall coefficient of determination was r^2=0.95. The values of coefficient of determination were higher for events that have clear time-points (e.g. tPNf, t2, t3) compared to events that have looser definitions and have higher inter-observer variation (e.g. tPNa and tM).



Annotationt	r^2
PNa	0.88
tPNf	0.99
t2	0.99
t3	0.97
t4	0.97
t5	0.97
t6	0.97
t7	0.97
t8	0.95
t9+	0.94
tM	0.92
tSB	0.91
tB	0.93
tEB	0.87
Total	0.95

Fig 2: Scatter plot of concordance between automated predicted annotation and expert's manual annotation of 15 morphokinetic events. The coefficient of determination (r^2) of each event and summary of the overall prediction of the model is showed in the table. tPNa- time to pronuclei appearance; tPNf- time to pronuclei fading; t(N)- time to cleavage of N discrete cells; tM- time to morula; tSB- time to start of blastulation; tEB- time to expanded blastocyst.

Comparison of morphokinetics to morphology

The database of KID embryos was divided into a training dataset composed of 6363 embryos and a testing dataset composed of 575 embryos (table 1).

Table 1: The distribution of KID embryos in training and test datasets

		Total	Total			Train		Test	
	Age		**N total**	**KID+**	**KID-**	**KID+**	**KID-**	**KID+**	**KID-**
Clinic 1	32.8±5.9	28485	4359	882	3477	776	3357	106	120
Clinic 2	33.0±7.0	1626	667	61	606	42	495	19	111
Clinic 3	34.6±5.8	3562	1739	198	1541	172	1379	26	162
Clinic 4	38.1±4.4	2888	173	108	65	88	54	20	11
Total	33.1±6.0	36561	6938	1249	5689	1078	5285	171	404

Note: KID, known implantation data; KID+, KID positive; KID-, KID negative

We compared the predictive ability of the three models (morphokinetic alone, morphology alone and combination of both) using area under the curve (AUC) calculations. We continuously evaluated AUC from 30 to 116 hours as shown in Figure 3A. The three models demonstrated increased predictive ability as time increased. The algorithm that assessed both morphokinetic and morphology was able to predict implantation more accurately with a higher AUC until 90 hours following insemination, than the algorithm that only considered the last frame. However, after 90 hours, the combined morphokinetic and morphology algorithm had a similar predictive ability when compared to the algorithm that only considered the last frame. Adding discarded embryos to the dataset significantly increased the AUCs of the three models (Fig. 3B).

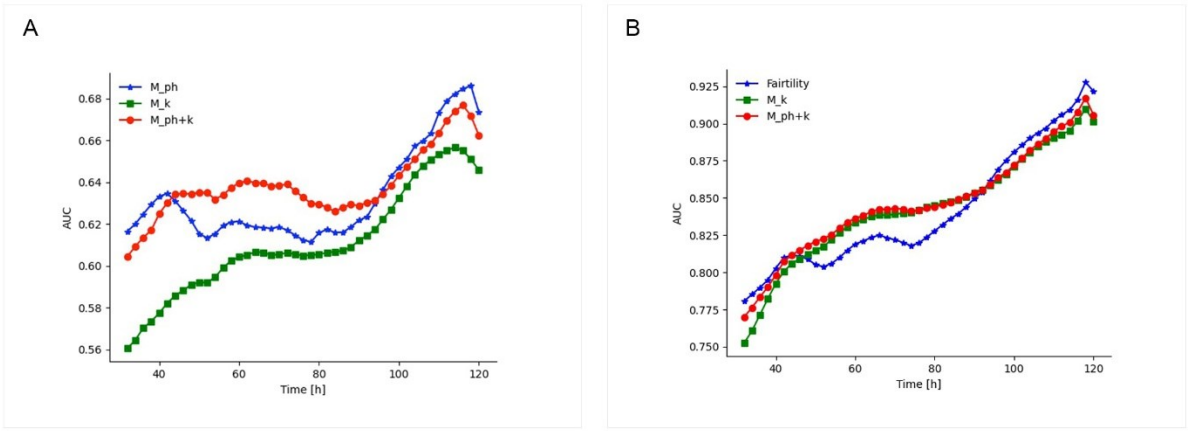


Fig 3: Comparison of the prediction (area under curve - AUC) of the three models (morphokinetic alone, morphology alone and combination of both) continuously from 30 to 116 hours of known implantation data embryos (A) or with discarded embryos (B). M_ph- morphology model, M_k-morphokinetic model, M_{ph+k}- combined morphology and morphokinetic model.

Discussion

In this study, we demonstrated an algorithm capable of automatically annotating key embryo developmental events. Incorporation of those estimated annotations into a CNN algorithm revealed a robust implantation prediction tool that continually improves every hour from 30 to 116 hours post insemination with a maximal AUC of 0.65 at 116 hours. However, analysis of the last frame of the embryo development video after the morula stage showed better prediction ability with AUC of 0.68 at 116 hours. When both algorithms are combined, we discovered that the morphokinetics annotation added value to morphology until the start of blastulation (around 90 hours). After that stage, the combined tool was as effective as the morphology alone algorithm with AUC of 0.68 at 116 hours.

Embryo evaluations involve two tasks: implantation prediction and embryo ranking. Most studies incorporate patient age into the algorithm which yield improved predictive ability [24, 25]. However, age has such a significant effect that it may mask the effects of morphology and morphokinetic analysis when predicting embryo implantation potential. As a result, incorporation of an algorithm that includes age into the clinical setting may be a disadvantage because in practice, the task is to rank embryos for a specific patient with a constant age. Therefore, in this study we wanted to isolate morphology and morphokinetic assessments and understand their contributions to implantation prediction. Thus, we have excluded age from the combined morphology and morphokinetic algorithm with the trade-off of having lower AUCs.

For the three models, morphokinetic, morphology, and combined morphokinetic and morphology, we assembled the implantation prediction AUC results as continuous time-points. An interesting pattern observed in all three of the models was an increase in AUC values until the first hours of day three, followed by a plateau from t8 that lasted until the early compaction stage. For all three models, the AUC values increased at the start of the blastulation stage and was sustained throughout blastocyst progression; this effect was most prominent in the morphology algorithm. This pattern represents the inherent challenge of evaluating morphology and morphokinetics of the embryonic compaction stages due to the lack of objective landmarks for annotation. This finding is consistent with previous studies that demonstrated the ability of tSB [26, 27], tEB and tB [28, 29] to predict implantation and the limitations of using morphokinetic parameters associated with the compaction stage. Moreover, these findings align with the scarce publication on morula morphology and implantation prediction [30]. To the best of our knowledge, this is the first study to show implantation prediction potential continuously by time (in hours).

A review by Reignier *et al*. reported the predictive value of morphokinetic parameters for embryo ploidy status. This comprehensive study and a study by Zaninovic *et al*. [31] showed that though the morphokinetic parameters differed significantly between aneuploid and euploid embryos, there is insufficient evidence for routine TLM use for embryo ploidy assessment. Another study evaluating morphology and morphokinetics parameters and ploidy status found no correlation between early stage parameters (cleavage stage) and ploidy status but did find a correlation between blastocyst morphology and late morphokinetics (from start of blastulation) parameters and ploidy status. However, the overlap between euploid and

aneuploid parameters was too big to define a prediction model [32]. By evaluating the algorithms in a continuous way, we overcame the previously demonstrated hurdles associated with blastocyst stage morphokinetic evaluation. We were able to develop a robust prediction model capable of evaluating morphology and morphokinetic parameters at all stages of embryo development. We have demonstrated the strengths and weaknesses of the morphology only and morphokinetic only algorithms and have shown that to achieve the best predictive algorithm, future prediction tools should incorporate both parameters.

In summary, this is the first fully automated morphology and morphokinetics machine learning model that evaluates implantation prediction over time. We have demonstrated that predictive power increases as development time increases and evaluating both morphology and morphokinetic parameters increased predictive power.

References

1. D'Alton, M. and N. Breslin, *Management of multiple gestations.* Int J Gynaecol Obstet, 2020. **150**(1): p. 3-9.
2. Grady, R., et al., *Elective single embryo transfer and perinatal outcomes: a systematic review and meta-analysis.* Fertil Steril, 2012. **97**(2): p. 324-31.
3. Gardner, D.K. and W.B. Schoolcraft, *Culture and transfer of human blastocysts.* Curr Opin Obstet Gynecol, 1999. **11**(3): p. 307-11.
4. Balaban, B., K. Yakin, and B. Urman, *Randomized comparison of two different blastocyst grading systems.* Fertil Steril, 2006. **85**(3): p. 559-63.
5. Gardner, D.K., et al., *Single blastocyst transfer: a prospective randomized trial.* Fertil Steril, 2004. **81**(3): p. 551-5.
6. Storr, A., et al., *Inter-observer and intra-observer agreement between embryologists during selection of a single Day 5 embryo for transfer: a multicenter study.* Hum Reprod, 2017. **32**(2): p. 307-314.
7. Castello, D., et al., *How much have we learned from time-lapse in clinical IVF?* Mol Hum Reprod, 2016. **22**(10): p. 719-727.
8. Zhan, Q., et al., *Direct Unequal Cleavages: Embryo Developmental Competence, Genetic Constitution and Clinical Outcome.* PLoS One, 2016. **11**(12): p. e0166398.
9. Rubio, I., et al., *Clinical validation of embryo culture and selection by morphokinetic analysis: a randomized, controlled trial of the EmbryoScope.* Fertil Steril, 2014. **102**(5): p. 1287-1294 e5.
10. Fishel, S., et al., *Evolution of embryo selection for IVF from subjective morphology assessment to objective time-lapse algorithms improves chance of live birth.* Reprod Biomed Online, 2020. **40**(1): p. 61-70.
11. Kaser, D.J. and C. Racowsky, *Clinical outcomes following selection of human preimplantation embryos with time-lapse monitoring: a systematic review.* Hum Reprod Update, 2014. **20**(5): p. 617-31.
12. Milewski, R., et al., *Morphokinetic parameters as a source of information concerning embryo developmental and implantation potential.* Ginekol Pol, 2016. **87**(10): p. 677-684.
13. Reignier, A., et al., *Can time-lapse parameters predict embryo ploidy? A systematic review.* Reprod Biomed Online, 2018. **36**(4): p. 380-387.
14. Swain, J.E., *Could time-lapse embryo imaging reduce the need for biopsy and PGS?* J Assist Reprod Genet, 2013. **30**(8): p. 1081-90.
15. Chen, F., et al., *Selecting the embryo with the highest implantation potential using a data mining based prediction model.* Reprod Biol Endocrinol, 2016. **14**: p. 10.
16. Richardson, A., et al., *A clinically useful simplified blastocyst grading system.* Reprod Biomed Online, 2015. **31**(4): p. 523-30.
17. Tanaka, H., *Modeling the motor cortex: Optimality, recurrent neural networks, and spatial dynamics.* Neurosci Res, 2016. **104**: p. 64-71.
18. He, K., et al., *Mask R-CNN.* IEEE Trans Pattern Anal Mach Intell, 2020. **42**(2): p. 386-397.

19. Wahab, N., A. Khan, and Y.S. Lee, *Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images.* Microscopy (Oxf), 2019. **68**(3): p. 216-233.
20. Meseguer, M., et al., *The use of morphokinetics as a predictor of embryo implantation.* Hum Reprod, 2011. **26**(10): p. 2658-71.
21. Saeedi, P., et al., *Automatic Identification of Human Blastocyst Components via Texture.* IEEE Trans Biomed Eng, 2017. **64**(12): p. 2968-2978.
22. VerMilyea, M., et al., *Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF.* Hum Reprod, 2020. **35**(4): p. 770-784.
23. Ciray, H.N., et al., *Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group.* Hum Reprod, 2014. **29**(12): p. 2650-60.
24. Tran, D., et al., *Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer.* Hum Reprod, 2019. **34**(6): p. 1011-1018.
25. Khosravi, P., et al., *Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization.* NPJ Digit Med, 2019. **2**: p. 21.
26. Mizobe, Y., et al., *Synchrony of the first division as an index of the blastocyst formation rate during embryonic development.* Reprod Med Biol, 2018. **17**(1): p. 64-70.
27. Goodman, L.R., et al., *Does the addition of time-lapse morphokinetics in the selection of embryos for transfer improve pregnancy rates? A randomized controlled trial.* Fertil Steril, 2016. **105**(2): p. 275-85 e10.
28. Kirkegaard, K., et al., *Time-lapse parameters as predictors of blastocyst development and pregnancy outcome in embryos from good prognosis patients: a prospective cohort study.* Hum Reprod, 2013. **28**(10): p. 2643-51.
29. Chamayou, S., et al., *The use of morphokinetic parameters to select all embryos with full capacity to implant.* J Assist Reprod Genet, 2013. **30**(5): p. 703-10.
30. Gardner, D.K. and B. Balaban, *Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': is looking good still important?* Mol Hum Reprod, 2016. **22**(10): p. 704-718.
31. Zaninovic, N., M. Irani, and M. Meseguer, *Assessment of embryo morphology and developmental dynamics by time-lapse microscopy: is there a relation to implantation and ploidy?* Fertil Steril, 2017. **108**(5): p. 722-729.
32. Minasi, M.G., et al., *Correlation between aneuploidy, standard morphology evaluation and morphokinetic development in 1730 biopsied blastocysts: a consecutive case series study.* Hum Reprod, 2016. **31**(10): p. 2245-54.

Solving the "right" problems for effective machine learning driven in-vitro fertilization

Itay Erlich[1,2], Assaf Zaritsky[3]

[1]The Alexander Grass Center for Bioengineering, School of Computer Science and Engineering, Hebrew University of Jerusalem, Israel

[2]Fairtilty Ltd., Tel Aviv, Israel

[3]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

[*]Corresponding author: Assaf Zaritsky, assafza@bgu.ac.il

Abstract

Automated live embryo imaging has transformed in-vitro fertilization (IVF) into a data-intensive field. Unlike clinicians who rank embryos from the same IVF cycle cohort based on the embryos visual quality and determine how many embryos to transfer based on clinical factors, machine learning solutions usually combine these steps by optimizing for implantation prediction and using the same model for ranking the embryos within a cohort. Here we establish that this strategy can lead to sub-optimal selection of embryos. We reveal that despite enhancing implantation prediction, inclusion of clinical properties hampers ranking. Moreover, we find that ambiguous labels of failed implantations, due to either low quality embryos or poor clinical factors, confound both the optimal ranking and even implantation prediction. To overcome these limitations, we propose conceptual and practical steps to enhance machine-learning driven IVF solutions. These consist of separating the optimizing of implantation from ranking by focusing on visual properties for ranking, and reducing label ambiguity.

Introduction

One of the major challenges in achieving effective in-vitro fertilization (IVF) is selecting the specific embryos for implantation from a cohort of "sibling" embryos from the same IVF cycle [1, 2]. The decision as to how many and which embryos to implant back to the uterus is driven by the goal of achieving the birth of exactly one healthy baby [3, 4]. Toward this goal, clinicians score the visual properties of the developing embryos, alongside non-visual clinical parameters that are common to all the embryos from the same cohort, such as previous IVF outcomes, oocyte age and underweight/obesity to estimate the probability of a successful implantation and use these parameters to decide how many and which embryos to implant [2, 5, 6, 7, 8] (Fig. 1A). Traditionally, embryo grading and selection are subjectively performed in a two-step process where clinicians first rank cohort embryos based solely on their visual qualities, as a proxy for embryo implantation potential, and then use non-visual clinical properties as a proxy for the embryo-extrinsic factors involved in implantation, to decide how many embryos to transfer [9, 7]. Recent advances in automated imaging and computation have led to objective and systematic data-driven approaches for embryo evaluation, where machine learning models are trained to output a score that predicts implantation potential [10, 11, 12, 13, 14, 15, 16]. These scores are then used to rank the embryos in an IVF cycle cohort and assist the clinician in making decisions regarding which embryos should be transferred. This approach of first using massive datasets to train a classifier that predicts retrospective implantation outcomes, and then using its prediction to rank embryos for transfer is performed under the implicit premise that a score trained to predict implantation potential is suitable for ranking embryos from a specific cycle cohort.

Here, we challenge this axiom. There is a general consensus that non-visual clinical properties contribute to the prediction of implantation outcomes [17, 18, 6, 14], but that such information is not relevant for ranking embryos in the context of a single IVF cycle cohort that share the same non-visual clinical properties. This implies that the optimization task solved through machine learning in practice is different from the actual decision made in the clinic. Moreover, while successful implantation inherently implies that an embryo is of high implantation quality, a failed implantation can be caused by a defective embryo or be related to poor maternal clinical factors (such as oocyte age, uterus cavity and receptivity

of the endometrium). This creates uncertainty in the ground truth labels of failed implanted embryos that may hamper the ability of machine learning models to generalize well during training. It is not clear whether using machine learning models trained with clinical information can hamper the selection of the most promising embryos for implantation, and what is the magnitude of the errors caused by the ambiguous labels of failed implantations.

We hypothesized that predicting implantation is not the optimal machine learning strategy for solving the embryo-ranking problem because features derived from clinical information shared among embryos from the same cohort are not relevant for the ranking task and because ambiguous labels confound the optimal ranking. We confirmed this hypothesis by comprehensively assessing the contribution of visual embryo properties and clinical factors, and by evaluating several criteria for training data selection with less ambiguous labels. Finally, we quantified the extent of these non-optimal strategies and propose conceptual and practical setups to overcome these limitations by training and applying separate classifiers for the different tasks of predicting embryo implantation potential and ranking.

Results

Decoupling embryo scoring and ranking using embryo morphology, morphokinetics and oocyte age

One important application of machine learning in IVF clinics is the prediction of successful implantation. This task involves training a machine learning classifier using a large number of embryos that either succeeded (KID-positive, KID-P) or failed (KID-negative, KID-N) to implant back to the uterus (Methods). During classifier training, each embryo is represented by its visual properties and non-visual clinical properties that are provided together with the implantation outcome as the target label (Fig. 1A). The trained classifier receives as input new embryos and predicts a score corresponding to their implantation potential; i.e., the probability for successful implantation. The visual embryo properties include the appearance of the embryo (termed morphology) and the time it takes to reach predefined stages in development (termed morphokinetics). The non-visual clinical properties include oocyte age, history of previous IVF treatments, and any other property that is shared across all the embryos from the same cohort and thus cannot be directly extracted from the specific embryo's image. The embryo-specific visual properties reflect the intrinsic embryo's potential for implantation, whereas the cohort-specific non-visual clinical properties correspond to the embryo-extrinsic parameters. Successful implantation is heavily dependent on both the embryo-specific visual appearance and on the cohort-specific non-visual clinical factors.

We sought to systematically assess the contribution of non-visual clinical properties to embryo implantation by comparing machine learning models that rely on visual embryo features with or without non-visual clinical properties. Our data were composed of time-lapse videos collected from 4 clinics: 7904 IVF cycles that included 1314 KID-P and 6485 KID-N for Train, and 380 KID-P and 637 KID-N for Test (Methods, Table S1). The non-visual clinical property chosen was the oocyte age, shared among all embryos from the same IVF cycle cohort, and is probably the most common embryo-extrinsic property used for the embryo implantation prediction task [17]. First, we validated that the KID-P embryos were associated with younger oocytes (Fig. 1B). Next, we preprocessed the images (Fig. S1) using the morphology-based classifier described in a companion paper

and in the Methods (Fig. S2A). This classifier provides a continuous score reflecting the implantation score for each embryo over time with or without the oocyte age as additional input to the embryo image (Methods, Fig. S2C). As expected, the classifier that had access to oocyte age outperformed the classifier that relied on morphology alone over time (Fig. 1C). To further verify that the oocyte age contributes to visual properties we devised a new classifier that calculated a continuous implantation score from the embryo morphokinetic properties (Fig. S2B, Methods). The timing of a set of fifteen hallmark morphokinetic events were identified automatically by combining machine-learning and dynamic programming (Figs. S3-4). These morphokinetic features were used to train a classifier to predict the implantation potential of an embryo based on its developmental history (see Methods for full details). Including the oocyte age as an additional feature improved the morphokinetic performance over time (Fig. 1D). Although the morphology-based classifier was more accurate than the morphokinetic-based classifier, the inclusion of oocyte age flipped the classification performance in favor of the morphokinetic model (Fig. 1E). This result suggests that the residual contribution of oocyte age to morphokinetic features was greater than the contribution of oocyte age to morphological features in the context of the implantation classification task. Correlating the temporal derivative of the corresponding classifier pairs showed that the temporal trend of the classification performance was more similar between the classifier pair that was trained with morphological features (with and without the oocyte age) than the classifier trained with the morphokinetic features, thus providing further evidence that the common classification-related information between morphology features and oocyte age was higher than between morphokinetic features and oocyte age (Fig. 1F). Since oocyte age is common to all cohort embryos and thus irrelevant for the ranking task, the prediction of the higher performing classifier for the implantation potential is not necessarily aligned with the clinician's task of ranking "sibling" embryos originating from the same IVF cycle. Hence, although the classifier trained with both morphokinetic features and oocyte age outperformed the classifier trained with both morphological features and oocyte age (especially in early embryo development), embryo morphological features seem more appropriate for the task of ranking high quality embryos. Thus, beyond providing a quantitative assessment of the contribution of oocyte age to implantation prediction, these results raise the possibility that clinical properties that

are common across embryos from the same cohort contribute to the task of embryo scoring toward implantation but are not the optimal properties for the task of embryo ranking.

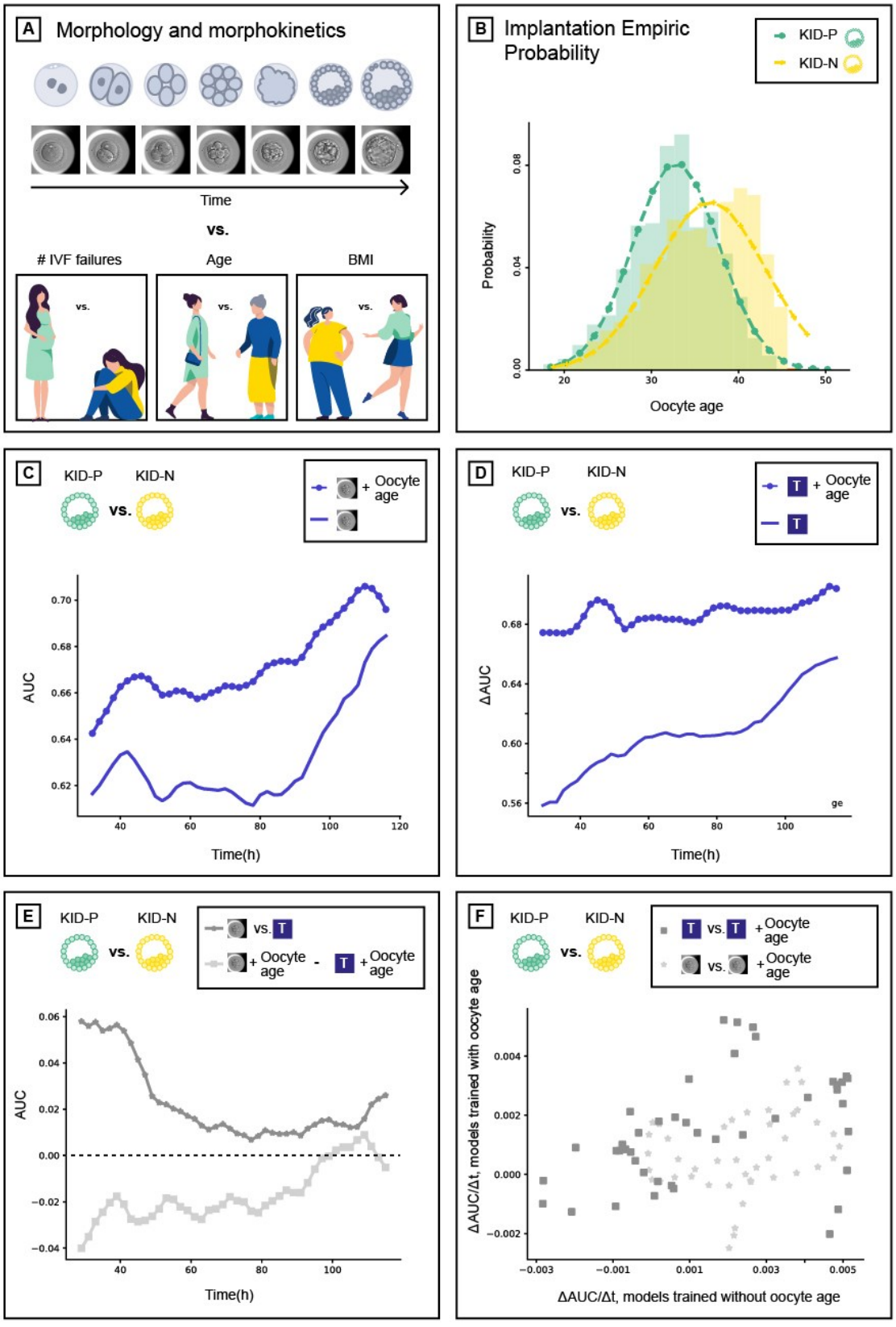

A
Morphology and morphokinetics
Time
vs.
IVF failures
Age
BMI
B
Implantation Empiric Probability
KID-P
KID-N
Probability
Oocyte age
C
KID-P
KID-N
vs.
+ Oocyte age
AUC
Time(h)
D
KID-P
KID-N
vs.
T
+ Oocyte age
ΔAUC
Time(h)
E
KID-P
KID-N
vs.
vs. T
+ Oocyte age - T + Oocyte age
AUC
Time(h)
F
KID-P
KID-N
vs.
T vs. T + Oocyte age
vs. + Oocyte age
ΔAUC/Δt, models trained with oocyte age
ΔAUC/Δt, models trained without oocyte age

Figure 1: Decoupling embryo scoring and ranking using intrinsic and extrinsic features. (**A**) Embryo implantation potential relies on the embryo's visual and clinical properties. (**B**) Distribution of implanted (KID-P) versus non-implanted (KID-N) embryos according to oocyte age. The average oocyte age of the KID-P (KID-N) embryos was 32.82 (36.74) with a standard deviation of 4.93 (6.33). N KID-P = 1694, N KID-N = 7122. (**C-D**) Comparison of the implantation prediction performance (AUC) of models trained with or without the oocyte age over time since fertilization. (**C**) Embryo morphological features (marked with an embryo image). (**D**) Embryo morphokinetic features (marked with "T"). (**E**) Comparing implantation prediction performance: the morphology-based classifier outperformed morphokinetic-based classifier but morphokinetic and oocyte age outperformed morphology and oocyte age. (**F**) Correlation between the temporal derivative of the morphology-based classification with and without oocyte age (Pearson correlation R = 0.43) versus correlation between the temporal derivative of the morphokinetic-based classification with and without oocyte age (Pearson correlation R = 0.27).

Optimizing features for prediction of embryo implantation is not an optimal strategy for the task of embryo ranking

Embryos in an IVF cycle are partitioned into three groups by the clinician based on their visual properties. These are composed of 1) transferred embryos - the most promising embryos for implantation (Video S1), 2) frozen embryos - other visually high-quality embryos that are kept for transfers in future IVF cycles and 3) discarded embryos - those selected to be excluded based on their poor visual appearance (Video S2-S4). Discarded embryos can be further partitioned into two subgroups: those that reached blastulation (termed *blastocyst-discarded*, Video S2) and those that did not reach blastulation (termed *underdeveloped-discarded*, Videos S3-4). These partitions define the expected ordered ranking of embryos based on their visual appearance (Fig. 2A).

A basic expectation of an embryo-ranking model is to prioritize KID-P over discarded embryos. To assess the performance of models trained to predict implantation results (KID-P versus KID-N) on the task of embryo ranking, we evaluated the average *time to live birth* (TTLB), the number of attempts for embryo transfer from an IVF cycle until a successful implantation was reached, where the transfer order was determined by the model's ranking based on its classification score [19]. For example, a TTLB value of 2 implies a cohort where the KID-P was ranked second; i.e., one discarded embryo had a higher classification score (Fig. 2B). The analysis included embryo cohorts that contained a successful implantation (KID-P) and multiple discarded embryos. While it is widely believed that transferring a discarded embryo will result in a failed implantation, the implantation

outcome for frozen embryos is not clear-cut; hence, they were not considered in this analysis. To avoid the bias of cohorts with different numbers of embryos we compared the average TTLB of sub-cohorts consisting of the same number of embryos sampled from the full cohorts (Methods).

Oocyte age contributed to improved implantation prediction (Fig. 1C-D) but this information is not relevant for ranking embryos from the same cohort that share the same oocyte age. To test the hypothesis that optimizing features for the KID-P versus KID-N classification task can deteriorate ranking accuracy, we compared the performance of two classifiers that were trained with morphological features with or without the oocyte age. We first ranked sub-cohorts of one KID-P and multiple discarded embryos (Methods). The average TTLB was at most 1.25 and the inclusion of oocyte age led to less efficient (i.e., higher TTLB) ranking, that required on average 1.7 attempts to achieve successful implantation (Fig. 2C). The same less efficient ranking with oocyte age was observed for sub-cohorts with one KID-P and multiple blastocyst-discarded embryos (Fig. 2D). As expected, ranking with blastocyst-discarded embryos constituted a more challenging task because the visual defects are less obvious with respect to embryos that did not reach blastulation. Similar trends were obtained for morphokinetic features with or without oocyte age (Fig. 2E-F). Altogether, these results suggest that selecting features to optimize implantation prediction is not the optimal strategy for embryo ranking, specifically when these features are clinical properties that are common to all embryos in a cohort.

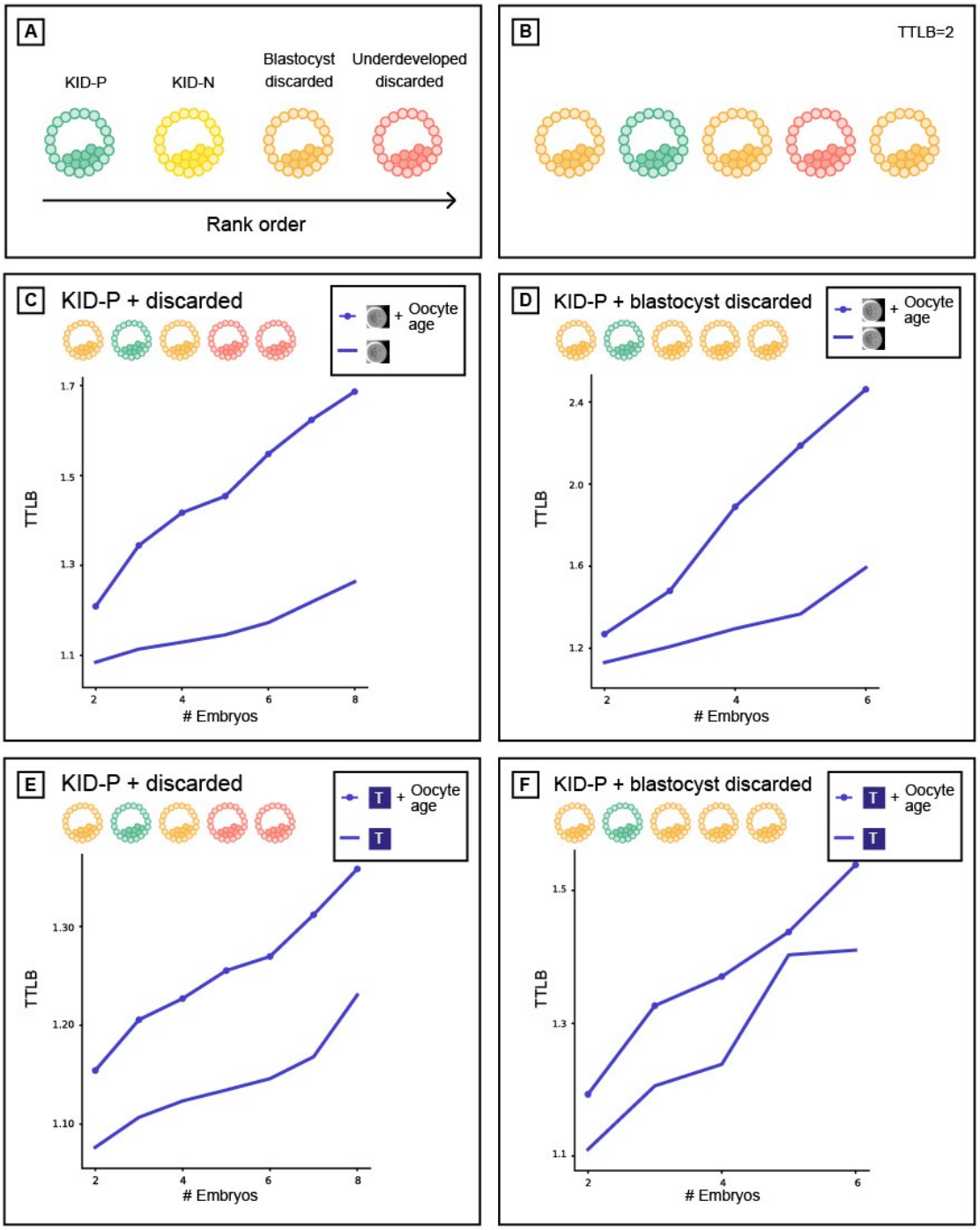


Figure 2: Using oocyte age as feature for improving implantation prediction reduced the accuracy of ranking embryos from the same cohort. (**A**) Embryo visual evaluation: KID-P > KID-N > Blastocyst-discarded > underdeveloped-discarded. KID-N and frozen embryos are not considered in embryo ranking evaluations because there is no clear distinction based on their visual properties. (**B**) Time To Live Birth (TTLB) for sub-cohorts of sibling embryos, as a

function of the number of embryos considered for each cohort. Each sibling sub-cohort consisted of one implanted embryo and multiple discarded embryos sampled from the full cohort. In the example here the number of embryos was 5 and the TTLB was 2 because the KID-P embryo was ranked second. (**C-D**) Morphology-based models (marked with an embryo image) with or without oocyte age. (**C**) Ranking KID-P and all discarded embryos. Number of cycles with n = 2-8 embryos: 201, 180, 163, 141, 115, 93, 67. Corresponding Wilcoxon signed rank tests on the null hypothesis that the two ranking schemes were drawn from the same matched distribution: <0.0001, <0.0001, <0.0001, <0.0001, <0.0001, <0.0001, 0.005. (**D**) Ranking KID-P and blastocyst-discarded embryos. Number of cycles with n = 2-6 embryos: 145, 98, 54, 32, 13. Corresponding Wilcoxon signed-rank tests: <0.0001, <0.0001, <0.0001, 0.001, 0.074. (**E-F**) Morphokinetic-based models (marked with "T") with or without oocyte age. Number of cycles according to panels C-D respectively. (**E**) Ranking KID-P and all discarded embryos. Wilcoxon signed-rank tests: 0.001, 0.001, 0.001, 0.001, 0.004, 0.004, 0.039. (**F**) Ranking KID-P and blastocyst-discarded embryos. Wilcoxon signed-rank tests: 0.004, 0.004, 0.071, 0.157, 0.655.

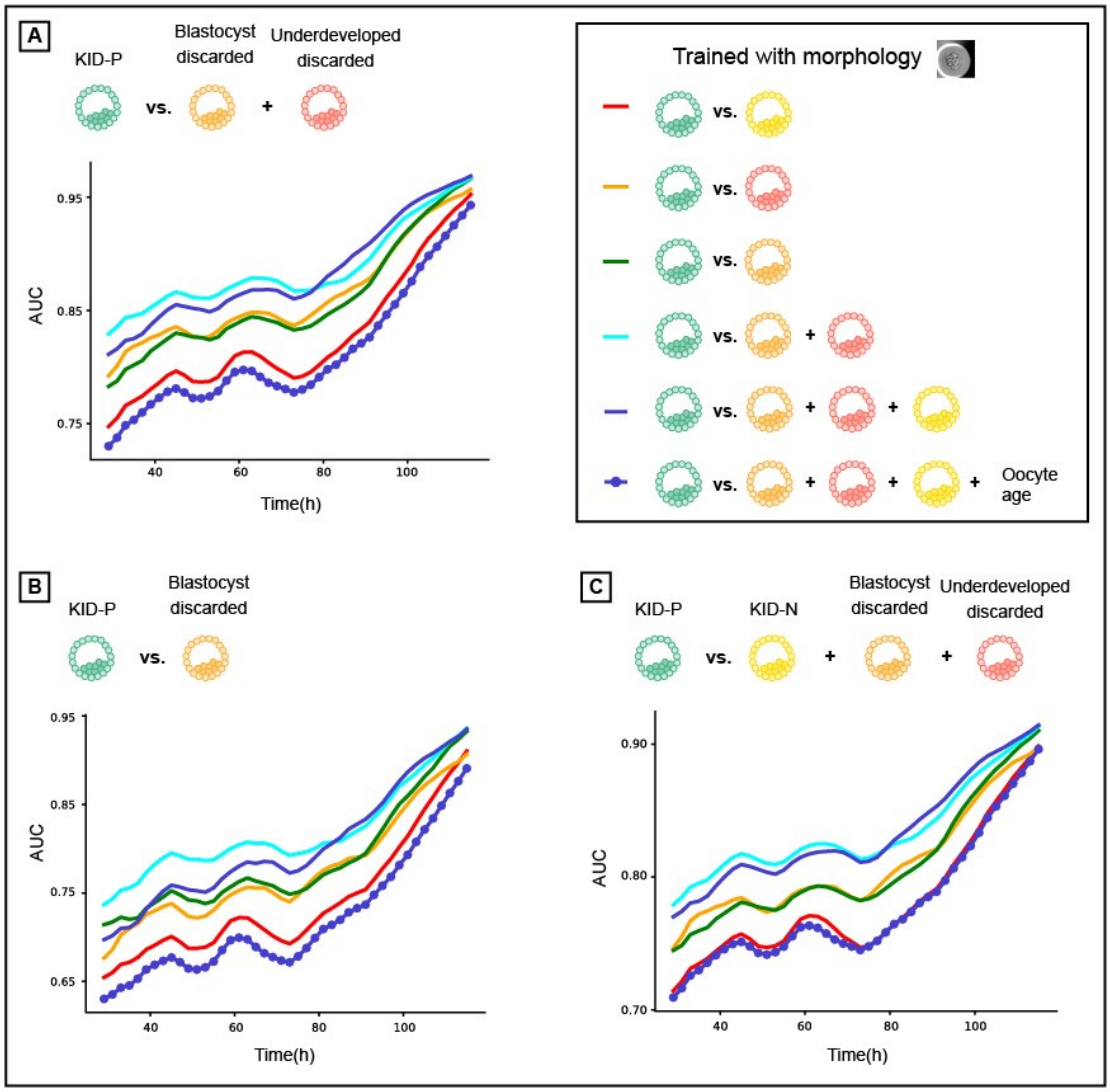
KID-N embryos are not the ideal negative labels for the task of embryo ranking

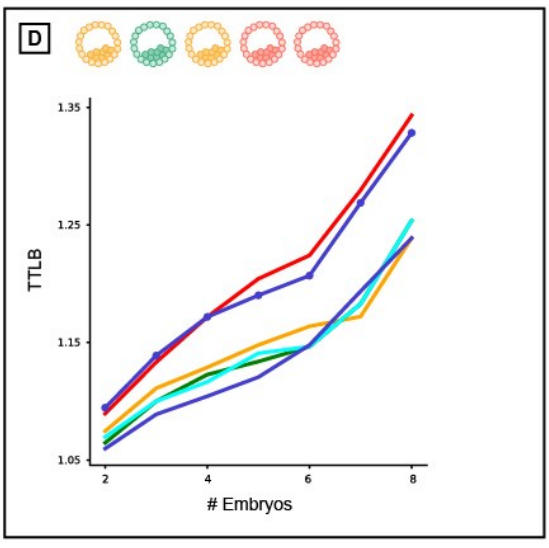
After demonstrating that the features selected for their performance at implantation prediction were not optimal for embryo ranking we asked how different training data would affect the models' performance when discriminating high quality from poor quality embryos. We defined a set of classification tasks that consisted of discriminating KID-P embryos from different subsets of discarded and/or unsuccessful implanted embryos. First, discriminating KID-P from all discarded embryos (Fig. 3A). This is an easy classification task because discarded embryos are characterized by evident visual defects. The second classification task involved discriminating KID-P from blastocyst-discarded embryos (Fig. 3B). The third involved discriminating KID-P from a subset that included both the KID-N as well as the discarded embryos (Fig. 3C). This latter setting is the one most similar to the true embryo ranking tasks in the clinic. For each of these embryo subsets we trained six morphology-based classifiers (Table S2) and their classification performance was compared on different classification tasks (Table S3, Fig. 3A-C).

Classifiers trained to predict implantation (KID-P versus KID-N) were less accurate than the classifiers that were trained with discarded embryos on all tasks (Fig. 3A-C). Inclusion of the oocyte age deteriorated the performance to the levels of the classifiers that were trained to predict implantation (Fig. 3A-C). Direct evaluation of time to live birth established that models trained to predict embryo implantation were not necessarily

optimal for the task of embryo ranking (Fig. 3D-E).

The KID-P versus KID-N labeled training data was severely limited in size because only a small subset of all embryos were selected for transfer and thus discarded embryos constitute the vast majority of embryos in IVF clinic in our study (Table S1) and in general [13, 20]. To assess the contribution of the extended volume of training data we limited the amount of negative labeled training data across the different categories of discarded embryos to align with the number of KID-N embryos (Table S2) and demonstrated that classifiers trained with discarded embryos outperformed the classifier trained to predict implantation even without extra training data (Fig. 3A-E, compare orange and green versus red). Additional training data further enhanced the classifiers' performance (Fig. S5). Similar results were obtained when using morphokinetic features (Figs. S6-7). Altogether, these results suggest that the elementary requirement of distinguishing poor from high quality embryos broke more frequently in the KID-P/KID-N trained classifiers, suggesting that KID-P/KID-N are not the optimal training data for the task of embryo ranking.


A
KID-P
Blastocyst discarded
Underdeveloped discarded
vs.
+
AUC
Time(h)
Trained with morphology
Oocyte age
B
KID-P
Blastocyst discarded
C
KID-P
KID-N
Blastocyst discarded
Underdeveloped discarded


D
TTLB
Embryos

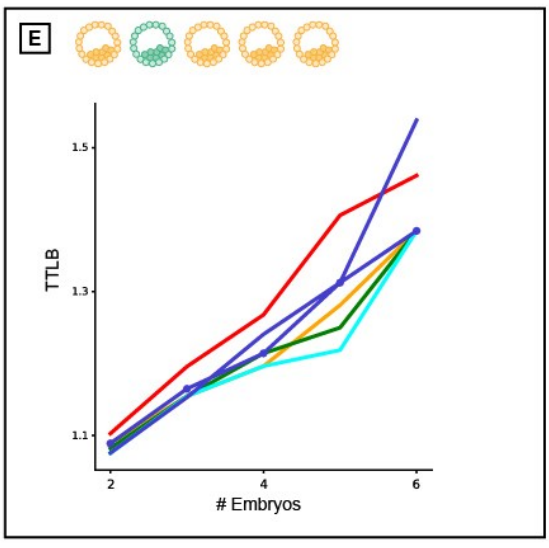

E
TTLB
Embryos

Figure 3: KID-N embryos are not the ideal negative label for solving the task of embryo ranking. (**A-C**) Prediction accuracy (AUC) over time since fertilization (higher scores reflect better performance). (**D-E**) Time To Live Birth (TTLB) for cohorts of sibling embryos, as a function of the number of embryos in each cohort (lower scores reflect better performance). Each siblings sub-cohort consisted of one implanted embryo and multiple discarded embryos from the same cohort. Number of cycles corresponding to those reported in Fig. 2. (**A-E**) All models were trained with morphological features. (**A**) KID-P versus discarded. (**B**) KID-P versus blastocyst-discarded. (**C**) KID-P versus KID-N and discarded. (**D**) Ranking KID-P and all discarded embryos. (**E**) Ranking KID-P and blastocyst-discarded embryos.

Tradeoffs between training labels ambiguity and specificity in predicting implantation and ranking

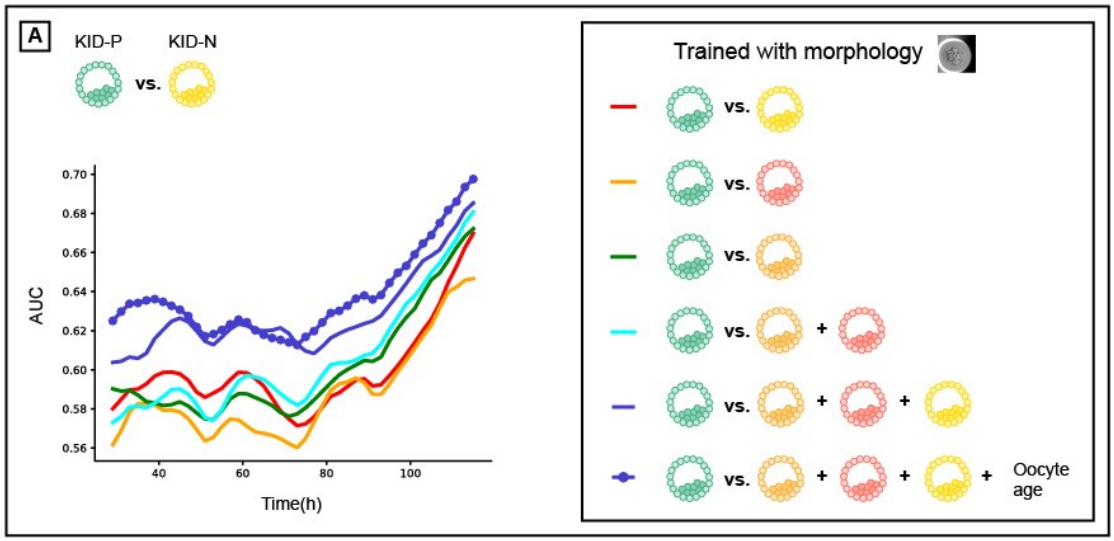
In the clinic, embryo ranking within an IVF cycle is exclusively determined by assessing the embryos' visual properties. In contrast, in the machine-learning solution, visual embryo properties only provide partial information in the context of implantation prediction. Failed implantations can be caused either by defected embryos or by embryo-extrinsic maternal clinical properties leading to non-receptive endometrium. This limitation translated to ambiguous (noisy) KID-N labels that can mislead the classifier during model training. The confidence in discarded embryos annotation, in the context of successful or failed implantation is high because the criterion of being discarded is inherently visual and thus is not/less affected by external factors.

To assess the contribution of training data volume and confidence in labels we turned again to the task of predicting implantation. We found that including the discarded embryos as negative labels in addition to KID-N during training, enhanced the classifier's performance in the KID-P versus KID-N classification task (Fig. 4A). Similar results were achieved when using morphokinetic features (Fig. S8). These results imply that despite the fact that the KID-P/KID-N classifier was trained for the most visually challenging task, the availability of extensive data and reliable labels improved classification performance.

The contribution of training label specificity to a particular classification task was demonstrated by the superiority of the model trained with the KID-N and the discarded embryos over the model trained with discarded embryos alone (Fig. 4A – blue versus cyan correspondingly, both have comparable volume of negative training labels). This result is strengthen by the lack of a similar trend in other classification tasks where the addition of

KID-N embryos to the training set did not contribute much to the performance (Fig. 3). These data demonstrate the tradeoff between label ambiguity and specificity and reveal that the combination of reduced ambiguity (but reduced specificity) in the discarded embryo labels along with increased specificity (but increased ambiguity) in the KID-N labels led to the most accurate implantation predictions.

To identify the breaking point between label ambiguity and specificity we simulated a gradual deterioration in the negative labels by flipping increasing fractions of positive labels to negative labels. This mimics the ambiguity of KID-N labels, where some viable embryos fail to implant for extrinsic maternal reasons and are thus mistakenly labeled as negatives. We evaluated the optimization problem of training models to discriminate KID-P from discarded embryos because of the higher confidence in the discarded embryos' negative label. This analysis revealed that flipping 10% of the discarded embryo labels was sufficient to drop the performance of the KID-P/discarded trained classifier below that of the KID-P/KID-N-trained classifier for the tasks of predicting implantation potential and ranking, thus underscoring the fact that small fractions of erroneous labels can cause major deteriorations in the model's performance (Fig. 4B-D). Altogether, these results highlight the interplay between label ambiguity, label specificity to a particular classification task and data volume during training in the context of IVF (Fig. 5).


A
KID-P
KID-N
vs.
AUC
Time(h)
Trained with morphology
vs.
vs.
vs.
vs. +
vs. + +
vs. + + +
Oocyte age

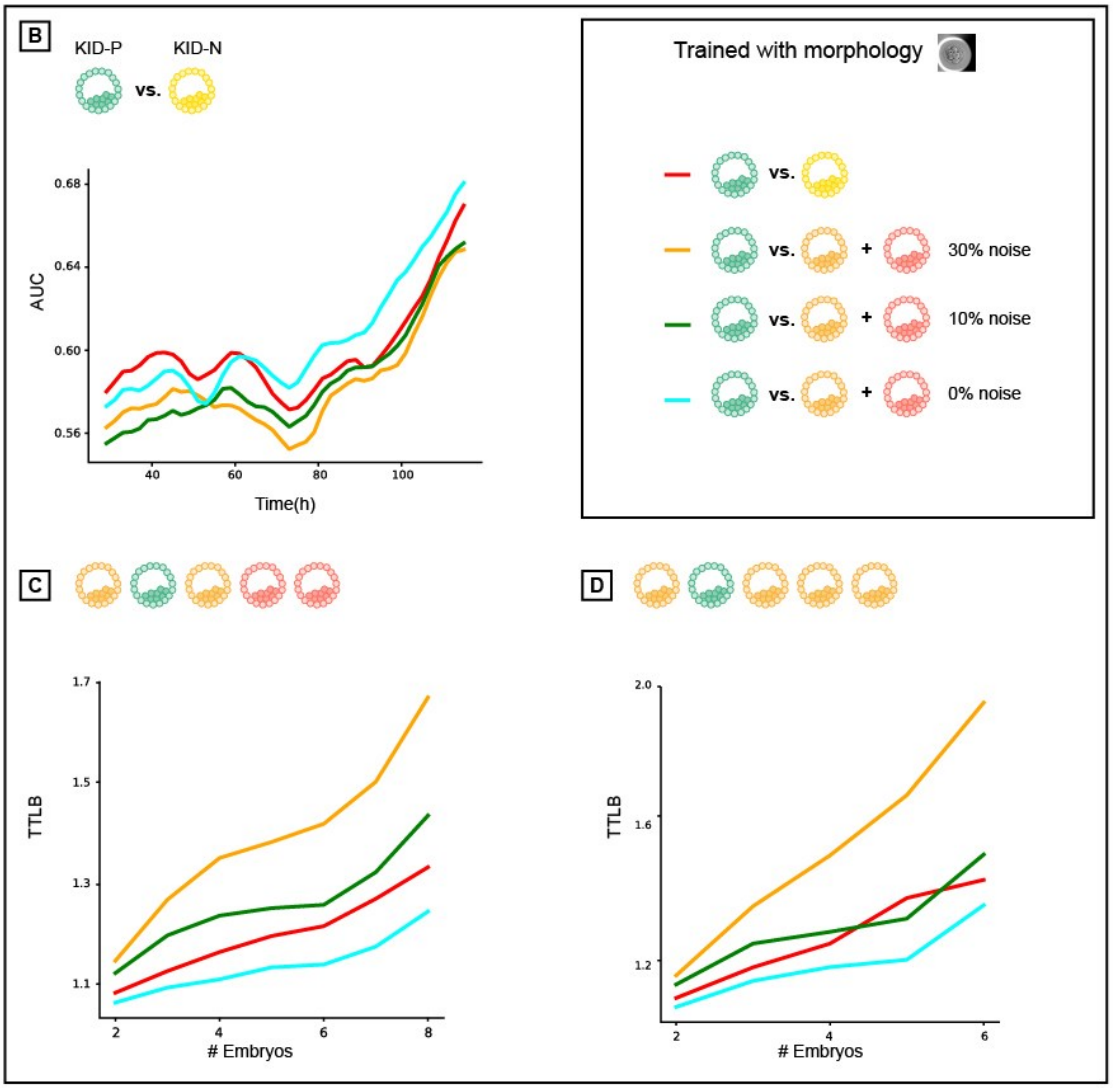

B
KID-P
KID-N
vs.
AUC
Time(h)
Trained with morphology
vs.
vs. + 30% noise
vs. + 10% noise
vs. + 0% noise
C
TTLB
Embryos
D
TTLB
Embryos

Figure 4: Analysis of implantation prediction and ranking performance as a function of negative label ambiguity and label specificity. All classifiers were trained using morphological features. (**A**) AUC of implantation prediction (KID-p versus KID-N) over time. Comparison of models trained to discriminate KID-P from different sets of negative labels: KID-N, discarded, blastocyst discarded, underdeveloped discarded, and KID-N+discarded with or without oocyte age. (**B**) AUC of implantation prediction (KID-p versus KID-N) over time for models trained to discriminate between KID-P and discarded embryos with increasing fractions of label ambiguity - flipping "KID-P" to "discarded" labels (see Methods). (**C-D**) Time To Live Birth (TTLB) for models trained in panel B. (**C**) KID-P versus all discarded. (**D**) KID-P versus blastocyst-discarded.

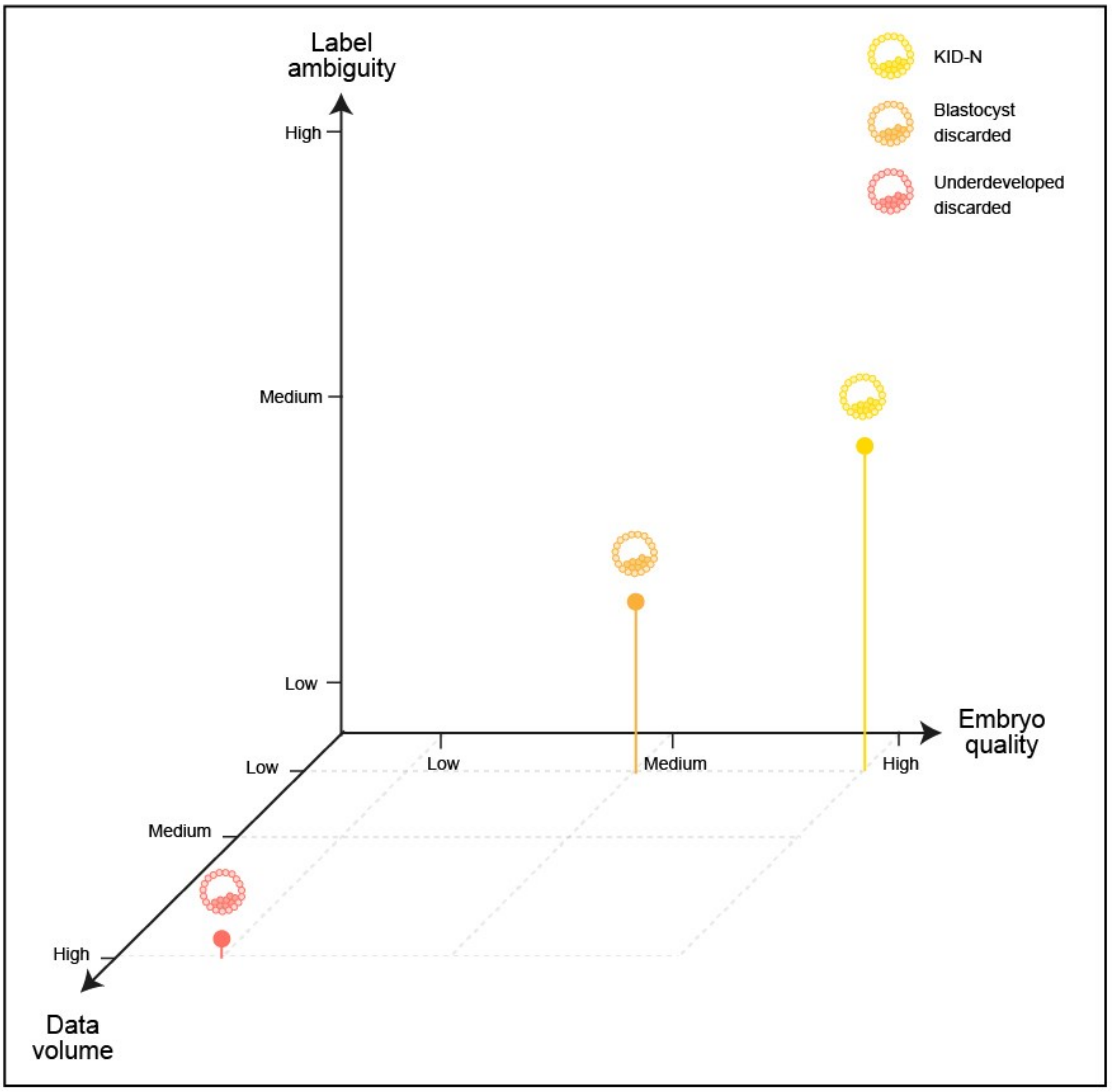


Figure 5: Trade-offs between data volume, label ambiguity and specificity to a particular classification task of the negative labels. In our case the specificity was aligned with the embryo quality (KID-P > blastocyst-discarded > underdeveloped-discarded, Fig. 2A). Data volume of underdeveloped-discarded was the highest (Table S1). Label ambiguity was high for KID-N. The positive labels were always KID-P.

Discussion

There is a consensus that clinical properties that are shared across all embryos within a cohort, such as oocyte age, play a key role in determining embryo implantation potential but have no effect on embryo ranking within a cohort. It is also well-accepted that some KID-N embryos are functional embryos that failed implantation due to embryo- extrinsic clinical factors. In this study we took the next conceptual step by demonstrating that training classifiers to predict implantation can lead to non-optimal transfer decisions as evident by the higher probabilities of preferring obviously visually defective embryos over successfully implanted embryos.

These results raise concerns as to the current common pipeline for machine-learning driven IVF that directly imply several practical recommendations. First, training models for embryo ranking should focus exclusively on embryo intrinsic features. These may include engineered features that measure specific embryo properties such as the appearance of pronuclei and fading timing (tPNa, tPNf), the number of pronuclei, pronuclei shape, symmetry [21, 22, 23] blastocyst expanded diameter and trophectoderm cell cycle length [24], temporal events [25, 26] and/or training deep neural networks on the raw images [13] while avoiding embryo- extrinsic clinical factors that are shared by the other sibling embryos within the cohort. Second, the ambiguity in the KID-N labels makes implantation prediction a sub-optimal optimization problem both in model training as well as in model evaluation. Thus training with less ambiguous negative labels, such as discarded embryos, can enhance ranking accuracy (and perhaps even implantation prediction). Training with less ambiguous negative labels can either be implemented by including discarded embryos in the training binary classification tasks, as demonstrated here, and/or by defining optimization problems that are more specific for ranking [27]. The current data suggest that this approach can be beneficial using discarded embryos as negative labels even though discarded embryos are commonly thought as trivially discriminated from high quality embryos (low label specificity). Beyond better reflecting the clinician's decision that involves the evaluation of all the embryos in an IVF cycle, another benefit of considering discarded embryos lies in the magnitude of data available for training that exceeded five-folds in the current dataset, and bypasses the biased selection of transferred embryos.

Hence avoiding/reducing the confounding effects of embryo- extrinsic clinical factors and ambiguous labels, when solving a ranking problem, as well as using all the available embryos to increase training data, are key practical concepts to consider when devising machine learning solutions for IVF. We hope that the awareness raised by our findings will bring back the traditional two-step process of ranking embryos within a cohort and determining how many embryos to transfer based on embryo scoring that approximates implantation probabilities, where each of these tasks is optimized independently under the appropriate assumptions.

Methods

Experiments

Data collection and ethics

The data were collected from four clinics: the Ein-Kerem, and Mt. Scopus campuses of the Hadassah Hebrew University Medical Center, the Soroka University Medical Center, and the NYU Langone Prelude Fertility Center. After fertilization, the embryos were incubated in EmbryoscopeTM (Vitrolife, Copenhagen, Denmark), a time-lapse incubation and imaging system that acquires seven-layers of z-stack images 15μm apart at each time point every 15 to 20 minutes, where time 0 is defined as the fertilization time. We used the central focal plane, which was usually the most focused, for further analysis, on a total of roughly 360-480 images over a 5-day period for a single embryo. Time-lapse sequences of embryos were collected between July 2014 and December 2019 for oocytes aged 23.6-43.6 years (Fig. 1B). Embryos that were discarded before 114 hours were excluded from further analysis. Human embryo image/video data collected from patients were used in this study with institutional review board approval from the Investigation Review Board of Hadassah Hebrew University Medical Center (IRB# HMO-006-20). Overall, our dataset contained 47162 recorded videos of embryos collected from 7904 patients with informed consent for research and publication, under an institutional review board approval for secondary research use. Table S1 details the data splits between train and test.

Annotation of embryo clinical quality

Once the embryos reached the desired developmental stage, single or more embryos from a cohort were transferred. For those embryos that were transferred, implantation tagging was defined by one of the three possible Known Implantation Data (KID) tags: (i) KID-positive (KID-P): when all transferred embryos were successfully implanted (Video S1); (ii) KID-negative (KID-N): when none of the transferred embryos were successfully implanted; (iii) KID-unknown (KID-UK): when the outcome of each transferred embryo was uncertain. KID-UK embryos were excluded from further analysis due to their ambiguous outcomes. Non-transferred embryos were partitioned into two groups: (1) Discarded – embryos that were excluded by the clinician because of defective morphology:

poor morphology [8] or morophkinetics (i.e., slow or halted development). The discarded embryos were further partitioned into two subgroups: "blastocyst-discarded" (Video S2), embryos that reached blastulation, and "underdeveloped-discarded" (Videos S3-4), that did not reach blastulation. Based on these selection criteria we assume that discarded embryos fail to implant, and thus used these discarded embryos as a non-ambiguous label to assess the model's performance when ranking KID-P over discarded embryos. (2) Frozen – visually proper embryos that were not selected for transfer because other sibling embryos from the same cohort were ranked higher. These embryos were frozen for a future transfer but were not considered for further analysis because their implantation potential was unknown. These partitions defined the embryo clinical ranking of implantation potential (Fig. 2A).

Analysis

Cropping and resizing embryo images

A simple, low-cost, Convolutional Neural Network (CNN) was trained to segment and localize the embryo. The time-lapse incubation produces embryo images of 500x500 pixels, at several focal planes (typically seven). Since the embryo captures nearly the whole image in this case, the segmentation network can accept a downscaled version of the original file with a size of 128x128 pixels to enhance efficiency and speed (1 focal plane, randomly chosen at train, focal 0 micron at test) and outputs a pixel- wise binary classification segmentation mask of size 128x128. A U-NET architecture [28] was used, with 4 convolutional layers at each of the downlink/uplinks. Each layer was followed by a maxpool/upsample layer, ReLU activation function and batch normalization, where the number of features started from 8 and was doubled after each pooling. A diagram of the architecture and examples of the segmentation mask are shown in Fig. S1. Finally, the bounded embryo was cropped and resized to 256x256 pixels image I_{ci}^{t}. For notational simplicity, x_i^t denotes the cropped and resized image I_{ci}^{t} in the remainder of this work.

Training classifier to discriminate KID-P embryos from different subsets of lower quality embryos

The data included the KID-P, KID-N, and discarded embryos that were split into Train/Validation/Test datasets (Tables S1). Each cohort of sibling embryos was allocated to one of the Train/Validation/Test datasets. Transferred embryos in the training set had incubation periods ranging from 2 to 6 days, whereas transferred embryos in the test set were composed solely of embryos that reached incubation durations of 114 to 120 hours to enable comparable results between different time points on the exact same set. This design makes the KID-N test set even more challenging compared to the training negative set since some of the embryos in the training set were transferred earlier on day 3 but would eventually have been labelled inviable, discarded and not transferred if they were allowed to incubate until day 5.

The learning task was to classify an input x^t, where x is a feature representation of an embryo and t is time, to its ground-truth implantation binary outcome (detailed below). Features representations of an embryo included morphology (raw image), morphokinetic (timing of a predetermined set of embryo developmental stages) and each of these with the oocyte age as an additional feature. Details on each of these feature sets can be found below. The loss function used to train all three models was the binary cross entropy loss, a special case of equation (2), with K=2. The output of the classifier was a scalar $s^t = s_2^t - s_1^t$ that represent the implantation probability p_t:

$$p^t = \frac{e^{s^t}}{1+e^{s^t}}, \tag{9}$$

thus, all the models were of the form $F: x^t \rightarrow s^t$, where the feature representation x^t changed between models and s^t is the scalar prediction.

We trained twelve different classifiers. Six classifier where trained with morphology-based features and another six classifiers with morphokinetic-based features. Each classifier was trained to discriminate KID-P embryos from different sets of embryos (Table S2): (1) KID-P vs. KID-N; (2) KID-P vs. underdeveloped ; (3) KID-P versus blastocyst discarded; (4) KID-P vs. all discarded; (5) KID-P vs. KID-N and all discarded; (6) KID-P vs. KID-P vs. KID-N and all discarded learned with oocyte age as an additional feature.

The morphology classifier, denoted M_h, received as input a single raw image. The morphokinetic classifier, denoted M_k, received a 15-dimensional vector holding the timing

of specific morphokinetic events that were automatically extracted from the time-lapse stream. These two classifiers are described next.

Classification using morphological features

The morphology-based model M_h accepts single images as input and therefore exploits the embryo morphological information to predict implantation. The model was implemented using a CNN with a modified ResNet50 architecture [32]. To compensate for the drastic morphological changes throughout the embryo developmental process we used the ResNet50 base layers with multiple prediction heads $\{h_n\}_{n\in[1\ldots\Delta_t/2]}$, each corresponding to a temporal window of two hour. Each head consisted of a 128 channel dense layer, followed by a scalar output layer s^t. Fig. S2A depicts the network's architecture. The morphology-based classier was described in a separate manuscript that is currently under review and is attached to the submission.

Automating the timing identification of morphokinetic events with machine-learning and dynamic programming

The problem of estimating the timing of early morphokinetic events was studied recently up to four [29, 30], and up to eight [31] cells in the developing embryo, where all the later stages were considered as one state. We extended this scope to quantify both early and late morphokinetic events. Specifically, we automatically extracted timing of 15 morphokinetic events in the embryo development, i.e., the timing of specific morphological changes during embryo development identified from the time-lapse images (Fig. S3-4, Video S1). These morphokinetic features were then used to train classifiers (Fig. S2B).

More formally, the *morphokinetic state S* was defined by the following K = 15 morphokinetic events: Zygote, Pronuclei (PN) appearance and fading, 1Cell, 2Cell ,¦3Cell, 4Cells, 5Cells, 6Cells, 7Cells, 8Cells, 9^+Cells, Morula, Start Blastulation, Blastocyst and Expanded Blastocyst.

$$S = \{\mathrm{Z}, \mathrm{PN}, 1\mathrm{C}, 2\mathrm{C}, 3\mathrm{C}, 4\mathrm{C}, 5\mathrm{C}, 6\mathrm{C}, 7\mathrm{C}, 8\mathrm{C}, 9^+\mathrm{C}, \mathrm{M}, \mathrm{SB}, \mathrm{B}, \mathrm{EB}\} \quad (1)$$

We combined machine-learning and dynamic programming to automated the timing identification of each specific morphokinetic event. The learning task was defined as classifying an image acquired at time t, x^t, to its corresponding morphokinetic event $y^t \in$

$[1 \ldots K]$, where the ground-truth annotations were determined manually by an expert for model training and evaluation. The training data were a set of the form $\{\{x_i^t, y_i^t\}_{i\in[N]}, t \in \Delta t_i\}$, where N was the number of embryos, and Δt_i was the time interval of the $i\text{-}th$ embryo time-lapse incubation stream.

The morphokinetic model M_k is a multi-class classifier, trained using a CNN with a ResNet50 architecture [32] that accepts a single image x_i^t and predicts its morphokinetic event. x_i^t represents a cropped and resized image of the centered and resized embryo inside its well. The loss function was Categorical Cross Entropy loss [33] defined as

$$L = -\frac{1}{N}\sum_{i\in[N],t\in\Delta_t} \sum_{k\in K} y_{i,k}^t \log(p_k^t), \tag{2}$$

$$p_k^t = \frac{e^{s_k^t}}{\sum_{k\in K} e^{s_k^t}}, \tag{3}$$

where s_k^t denotes the $k\text{-}th$ output of M_k at time t. p_k^t is derived from s_k^t and represents the likelihood ($0 \leqslant p_k^t \leqslant 1$) of each image x^t to be in $k-th$ state. $y_{i,k}^t$ is the one-hot encoded ground-truth. Namely,

$$y_{i,k}^t = \begin{cases} 1, & y_i^t = k \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

We denote by T, the set of K-1 morphokinetic events, defined by the starting time of all states in equation (1), except the first zygote, which is trivially defined by the time of the first frame $x_i^0, \ \forall i \in [N]$. Namely,

$$T = \{t_{\mathrm{PNa}}, t_{\mathrm{PNf}}, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_{9+}, t_M, t_{\mathrm{SB}}, t_{\mathrm{B}}, t_{\mathrm{EB}}\} \tag{5}$$

where t_X denotes the starting time of the X-th event.

The K size vector p^t can be thought as a "soft annotation" per each frame. To estimate the sequence of morphokinetic events given the per-frame soft annotations, we used dynamic programming, where the updating step at time $t+1$ is given by

$$\begin{aligned} Q_k^{t+1} &= \{Q_{\tilde{k}}^t, k\} \\ \tilde{k} &= \mathrm{argmax}_{j\in\{k,k-1\}}\{q_j^t\} \end{aligned} \tag{6}$$

where Q_k^t is a vector of size t, with the morphokinetic path at each time point $t \in [1, t]$,

that ends at the k-*th* state at time t, and q_k^t denotes a scalar score associated with the path Q_k^t. Where Q_k^t denotes the optimal path of state k at time t, in the sense that Q_k^t has the highest score q_k^t out of all possible paths that end at the k-*th* state at time t.

Further, the updated score q_k^{t+1} is given by

$$q_k^{t+1} = q_{\tilde{k}}^t + p_k^{t+1} \tag{7}$$

Simply stated, at each time $t+1$, and for each state k, the optimal paths at time t of the states $\{k-1, k\}$ are compared. The best path of the k-*th* state at time $t+1$ is the path with the highest score at time t, concatenated with the state k at time $t+1$. The new path score is the sum of the score of the best path from the $\{k-1, k\}$ states and the score of the soft annotations score of the k-*th* state at time $t+1$. Finally, the estimated ME vector $\boldsymbol{m} = [m_1, m_2, \ldots . m_{k-1}]^T$ is obtained by travelling over the transitions between states in Q_k^t. Namely:

$$m_k = t, \text{if } Q_k^t = k \cap Q_k^{t-1} = k-1 \tag{8}$$

Fig. S3 demonstrates the soft annotations and dynamic programming outcomes for embryos with different developmental paths.

The accuracy of the annotation classification was measured by two measurements: (1) the average per frame accuracy, calculated as the average the 0-1 loss over frames, and (2) the confusion matrix of the morphokinetic events. Fig. S4 presents the confusion matrices of morphokinetic events for the morphokinetic classifier before and after applying the dynamic programming stage, where rows denote the ground truth annotations, and columns the multi-class prediction. The confusion rates were prominent only between adjacent events. The first and final events presented relatively low confusion rates, whereas the intermediate morphokinetic events were subject to greater confusion rates with adjacent events. For example, distinguishing between 5C and 6C is more difficult than between 1C and 3C or between the morula and start of blastulation. These results align with annotation errors by expert embryologists. One exception is a blastocyst embryo, which the model confused with expanded blastocyst. However, this can be explained by the relatively vague definition of expansion, which in turn led to an inter (and even intra) variability in annotating the exact timing of when expansion began. The confusion matrix for the

dynamic programming decision executed over the multi-class prediction exhibited a higher true positive rate for all events and improved average accuracy to 85.5% compared to 83.6% for the multi-class prediction. This is attributed to the consideration of the full developmental trajectory as a global optimization problem.

Classification using morphokinetic features

The morphokinetic-based model M_k accepts as input the estimated ME vector $\boldsymbol{m}$, as calculated at time t. To allow the network to recognize embryos that stopped developing the time t was embedded to input vector. Formally, the input for model $\boldsymbol{M_k}$ was the K-dimensional vector $\boldsymbol{m}^t$, defined as

$$m_k^t = \begin{cases} m_k, & m_k \leqslant t, k < K \\ \infty & m_k > t, k < K \\ t, & k = K \end{cases} \quad (10)$$

The classifier was implemented with a neural network that consisted of 6 consecutive dense layers. The first 5 layers with 256 channels, and the last hidden layer consisted of 128 channels. Each layer was followed by ReLU activation and batch normalization. The final output layer was the scalar prediction. Fig. S2B depicts the network's architecture.

Classification using oocyte age

Oocyte age was previously was reported to be highly correlated with implantation rates [18]. We included as an additional input to the morphology- and morphokinetic-based models, resulting in two new models denoted by $\boldsymbol{M_{ho}}, \boldsymbol{M_{ko}}$ respectively. The architecture of these models was identical to their original models with the addition of an embedded layer that accepted oocyte age, digitized by one-hot encoding to one of five possible age groups, and outputted 128 features that corresponded to the hot age group. This output was added to the last hidden layer of the original model, followed by a dense layer with 128 channels and a scalar output layer. Fig S2C depicts the network's architecture.

Evaluating classifier performance in predicting embryo implantation and ranking

We evaluated the trained models' performance with Area Under the Curve (AUC) measurement on four different binary classification tasks of discriminating KID-P embryos from different sets of embryos (Table S3): (1) KID-P vs. KID-N; (2) KID-P vs. all

discarded embryos; (3) KID-P vs. blastocyst discarded; (4) KID-P vs. KID-N and all discarded embryos. This latter setting is most resembling the true embryo ranking in the clinic.

To directly assess the task of embryo ranking we evaluated the average number of attempts for embryo transfer from an IVF cycle until a successful implantation was reached, where the transfer order was determined by the model's ranking based on its classification score. This measurement is called *time to live birth* (TTLB) [19]. This analysis included embryo cohorts that contained a successful implantation (KID-P) and multiple discarded embryos, where transferring a discarded embryo is assumed to result in a failed implantation. Thus, a basic requirement from a ranking model is to prioritize KID-P over discarded embryos. To avoid the bias of cohorts with different numbers of embryos we compared the average TTLB of sub-cohorts consisting of the same number of embryos sampled from the full cohorts. For a given number of embryos (N = 2-8) we created sub-cohorts by selecting the KID-P embryo and N-1 discarded embryos. We considered two scenarios: (1) calculating TTLB for sets of one KID-P and discarded embryos (blastocyst and underdeveloped); (2) calculating TTLB for sets of one KID-P and multiple blastocyst-discarded embryos (more challenging).

Funding and Acknowledgments

This research was supported by the Israel Council for Higher Education (CHE) via the Data Science Research Center, Ben-Gurion University of the Negev, Israel, by the Israel Science Foundation (grant No. 2516/21) and by the Welcome Leap Delta Tissue program (to AZ). We thank Maya Adani for designing and generating the figures. We thank Nadav Rappoport, Eran Eshed, Assaf Ben-Meir and Oded Rotem for critically reading the manuscript.

Author Contribution

IE conceived the study, developed analytic tools, and analyzed the data. IE and AZ interpreted the data and drafted the manuscript. AZ mentored IE.

Competing Financial Interests

IE is an employee at Fairtility LTD. AZ is collaborating with AIVF LTD on projects not related to this study.

References

[1] D. K. Gardner and B. Balaban, "Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': is looking good still important?," *MHR: Basic science of reproductive medicine,* 2016.

[2] Y. Tian, W. Wang, Y. Yin, W. Wang, F. Duan and S. Zhao, "Predicting pregnancy rate following multiple embryo transfers using algorithms developed through static image analysis," *Reproductive biomedicine online,* vol. 34, p. 473–479, 2017.

[3] J. M. R. Gerris, "Single embryo transfer and IVF/ICSI outcome: a balanced appraisal," *Human Reproduction Update,* vol. 11, p. 105–121, 2005.

[4] A. Pinborg, "IVF/ICSI twin pregnancies: risks and prevention," *Human reproduction update,* vol. 11, p. 575–593, 2005.

[5] L. van Loendersloot, S. Repping, P. M. M. Bossuyt, F. van der Veen and M. van Wely, "Prediction models in in vitro fertilization; where are we? A mini review," *Journal of advanced research,* vol. 5, p. 295–301, 2014.

[6] B. Raef and R. Ferdousi, "A review of machine learning approaches in assisted reproductive technologies," *Acta Informatica Medica,* vol. 27, p. 205, 2019.

[7] D. K. Gardner, M. Lane, J. Stevens, T. Schlenker and W. B. Schoolcraft, "Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer," *Fertility and sterility,* vol. 73, p. 1155–1158, 2000.

[8] J. Zhang, H. Liu, X. Mao, Q. Chen, Y. Fan, Y. Xiao, Y. Wang and Y. Kuang, "Effect of body mass index on pregnancy outcomes in a freeze-all policy: an analysis of 22,043 first autologous frozen-thawed embryo transfer cycles in China," *BMC medicine,* vol. 17, p. 1–9, 2019.

[9] A. Scientists, "The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting," *Human reproduction,* vol. 26, p. 1270–1283, 2011.

[10] R.-S. Guh, T.-C. J. Wu and S.-P. Weng, "Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes," *Expert Systems with Applications,* vol. 38, p. 4437–4449, 2011.

[11] G. Corani, C. Magli, A. Giusti, L. Gianaroli and L. M. Gambardella, "A Bayesian network model for predicting pregnancy after in vitro fertilization," *Computers in biology and medicine,* vol. 43, p. 1783–1792, 2013.

[12] P. Malinowski, R. Milewski, P. Ziniewicz, A. J. Milewska, J. Czerniecki and S. Wołczyński, "The use of data mining methods to Predict the Result of Infertility Treatment Using the IVF ET Method," *Studies in Logic, Grammar and Rhetoric,* vol. 39, p. 67–74, 2014.

[13] D. Tran, S. Cooke, P. J. Illingworth and D. K. Gardner, "Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer," *Human Reproduction,* vol. 34, p. 1011–1018, 2019.

[14] A. Chavez-Badiola, A. F.-S. Farias, G. Mendizabal-Ruiz, R. Garcia-Sanchez, A. J. Drakeley and J. P. Garcia-Sandoval, "Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning," *Scientific reports,* vol. 10, p. 1–6, 2020.

[15] M. R. Hassan, S. Al-Insaif, M. I. Hossain and J. Kamruzzaman, "A machine learning approach for prediction of pregnancy outcome following IVF treatment," *Neural computing and applications,* vol. 32, p. 2283–2297, 2020.

[16] M. VerMilyea, J. M. M. Hall, S. M. Diakiw, A. Johnston, T. Nguyen and D. Perugini, "Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF," *Hum Reprod,* vol. 35, p. 770â\text{..}{8}{4}, 2020.

[17] A. S. T. Lim and M. F. H. Tsakok, "Age-related decline in fertility: a link to degenerative oocytes?," *Fertility and Sterility,* vol. 68, p. 265–271, 1997.

[18] F. Chen, D. De Neubourg, S. Debrock, K. Peeraer, T. D'Hooghe and C. Spiessens, "Selecting the embryo with the highest implantation potential using a data mining based prediction model," *Reproductive Biology and Endocrinology,* vol. 14, p. 1–12, 2016.

[19] S. K. Sunkara, W. Zheng, T. D'Hooghe, S. Longobardi and J. Boivin, "Time as an outcome measure in fertility-related clinical studies: long-awaited," *Human Reproduction,* vol. 35, p. 1732–1739, 2020.

[20] P. Khosravi, E. Kazemi, Q. Zhan, J. E. Malmsten, M. Toschi, P. Zisimopoulos, A. Sigaras, S. Lavery, L. A. D. Cooper, C. Hickman and others, "Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization," *NPJ digital medicine,* vol. 2, p. 1–9, 2019.

[21] J. Aguilar, Y. Motato, M. J. Escribá, M. Ojeda, E. Muñoz and M. Meseguer, "The human first cell cycle: impact on implantation," *Reproductive biomedicine online,* vol. 28, p. 475–484, 2014.

[22] L. Scott, R. Alvero, M. Leondires and B. Miller, "The morphology of human pronuclear embryos is positively related to blastocyst development and implantation," *Human reproduction,* vol. 15, p. 2394–2403, 2000.

[23] M. Li, W. Zhao, W. Li, X. Zhao and J. Shi, "Prognostic value of three pro-nuclei (3PN) incidence in elective single blastocyst-stage embryo transfer," *International journal of clinical and experimental medicine,* vol. 8, p. 21699, 2015.

[24] L. Bori, E. Paya, L. Alegre, T. A. Viloria, J. A. Remohi, V. Naranjo and M. Meseguer, "Novel and conventional embryo parameters as input data for artificial neural networks: an artificial intelligence model applied for prediction of the implantation potential," *Fertility and Sterility,* vol. 114, p. 1232–1241, 2020.

[25] B. M. Petersen, M. Boel, M. Montag and D. K. Gardner, "Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3," *Human reproduction,* 2016.

[26] Y. Liu, V. Chapple, K. Feenan, P. Roberts and P. Matson, "Time-lapse deselection model for human day 3 in vitro fertilization embryos: the combination of qualitative and quantitative measures of embryo growth," *Fertility and sterility,* vol. 105, p. 656–662, 2016.

[27] A. Chavez-Badiola, A. Flores-Saiffe-Farías, G. Mendizabal-Ruiz, A. J. Drakeley and J. Cohen, "Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation," *Reproductive BioMedicine Online,* vol. 41, p. 585–593, 2020.

[28] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[29] Z. Liu, B. Huang, Y. Cui, Y. Xu, B. Zhang, L. Zhu, Y. Wang, L. Jin and D. Wu, "Multi-task deep learning with dynamic programming for embryo early development stage classification from time-lapse videos," *IEEE Access,* vol. 7, p. 122153–122163, 2019.

[30] N. H. Ng, J. McAuley, J. A. Gingold, N. Desai and Z. C. Lipton, "Predicting embryo morphokinetics in videos with late fusion nets & dynamic decoders," 2018.

[31] J. Malmsten, N. Zaninovic, Q. Zhan, Z. Rosenwaks and J. Shan, "Automated cell division classification in early mouse and human embryos using convolutional neural networks," *Neural Computing and Applications,* vol. 33, p. 2217–2228, 2021.

[32] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv preprint arXiv:1512.03385,* 2015.

[33] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 21, p. 238–238, 1959.

Supplementary figures

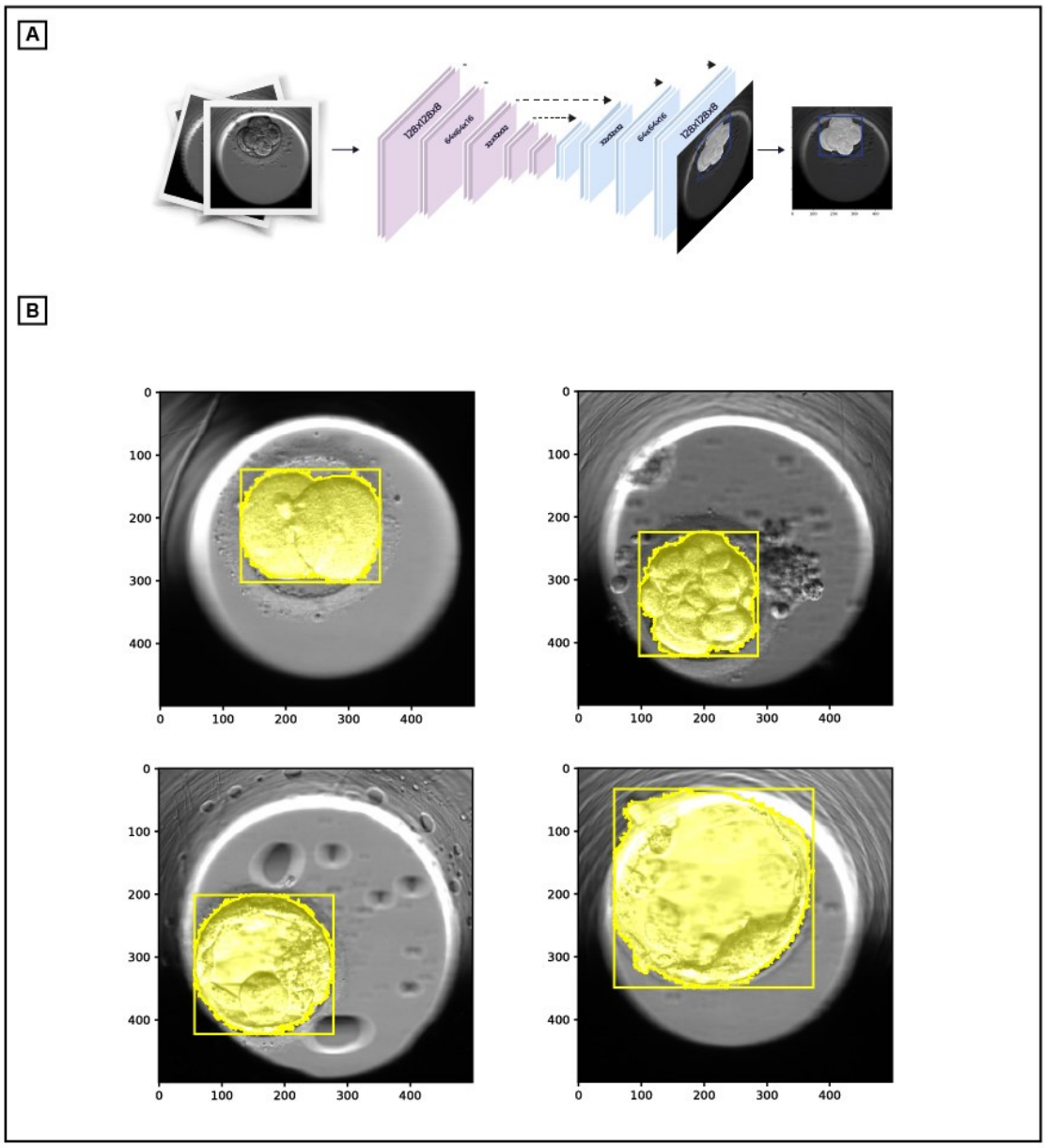


Figure S1: Embryo segmentation network. (**A**) U-NET architecture for embryo localization and segmentation. (**B**) Output masks for different developmental stage embryos. From upper left to lower right: 2 cells, 9+ cells, blastocyst, expanded blastocyst.

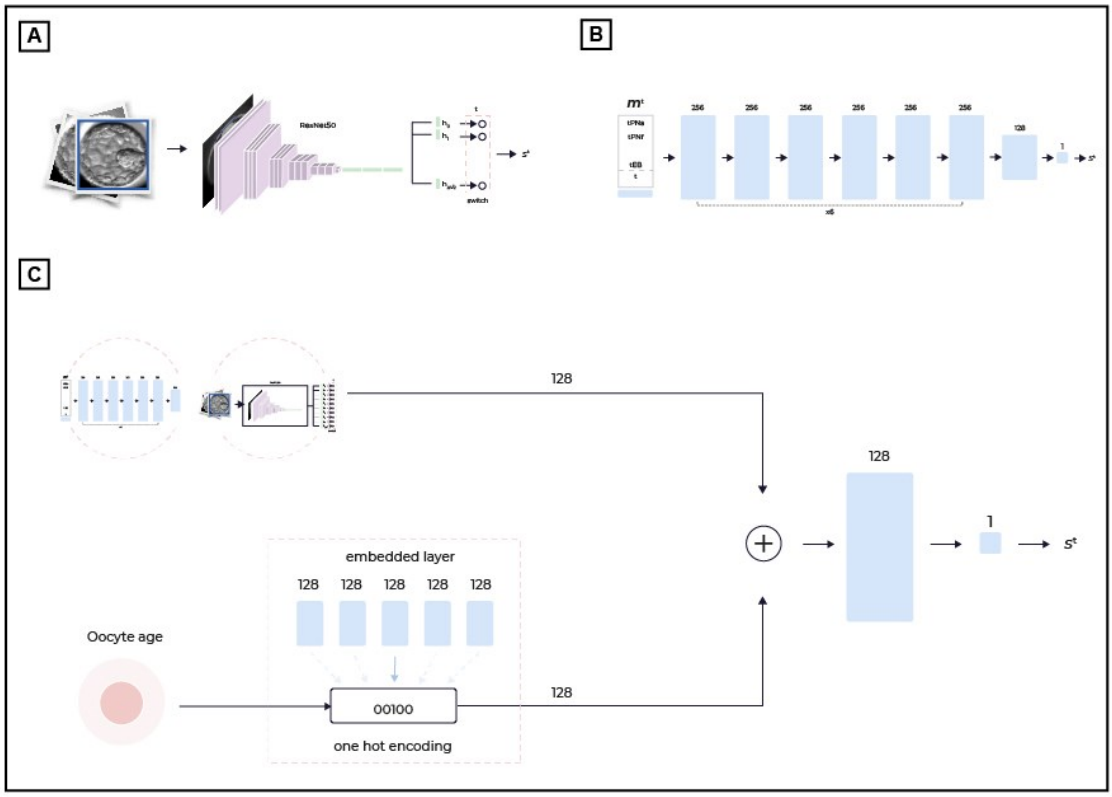


Figure S2: Implantation model architectures. (**A**) Morphological model M_h, 128x128 input image, fed into Resnet50 CNN with multiple prediction heads. Each head corresponds to a temporal window of 2 hours. (**B**) Morphokinetic model M_k, accepts vector $\boldsymbol{m}$ with the estimated starts of 15 morphokinetic events, and the time from fertilization. A simple neural network with eight dense layers was used. (**C**) The oocyte age was added as input to each model, resulting in models M_{ho}, M_{ko} respectively. The oocyte age was passed through an embedded layer with 128 output channels. The embedding layer digitized the oocyte age to one of 5 age bins, and outputted the corresponding 128 channels. The embedded output was then added to the output of each of the basis models. Two dense layers conclude the architecture.

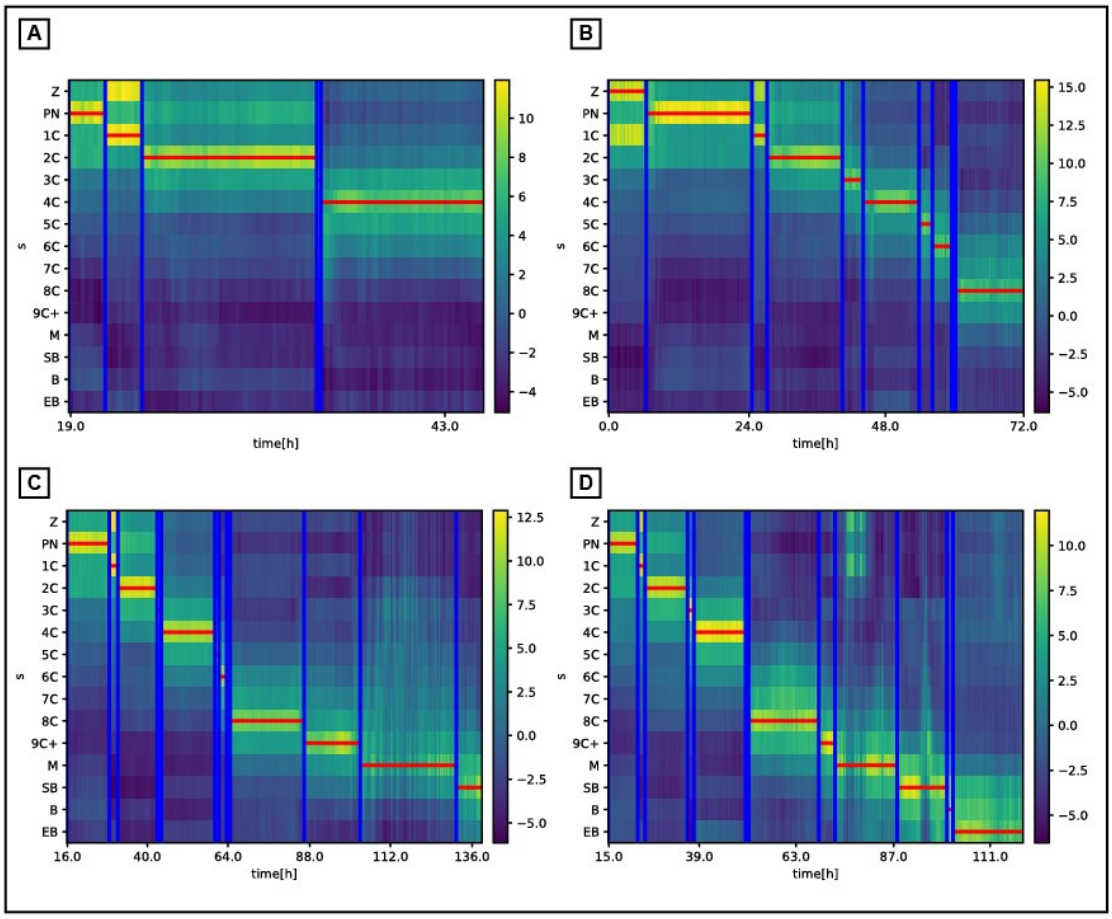


Figure S3: Predicted morphokinetic events vs. the ground truth (manual) annotations of different ending stage embryos: (**A**) 4Cells, (**B**) 8Cells, (**C**) Start Blastulation, (**D**) Expanded Blastocyst. The map represents the network multi-class prediction s^t of each frame vs. time. The horizontal lines are the dynamic programming output path Q_k^t, executed over the multi-class prediction map, and the vertical lines are the ground truth (manual) annotations. See Methods for full details.

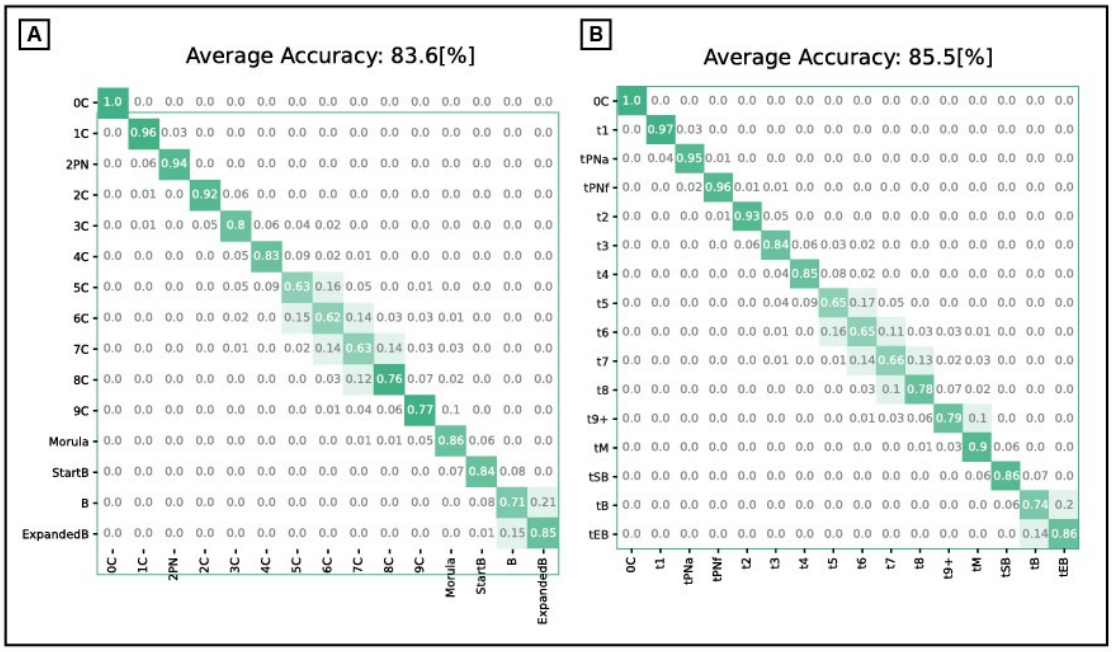


Figure S4: Confusion matrix of morphokinetic state classification. Rows denote the ground truth annotations, and columns the multi-class prediction. Shown are multi-class model predictions before (**A**), and after (**B**) dynamic programming.

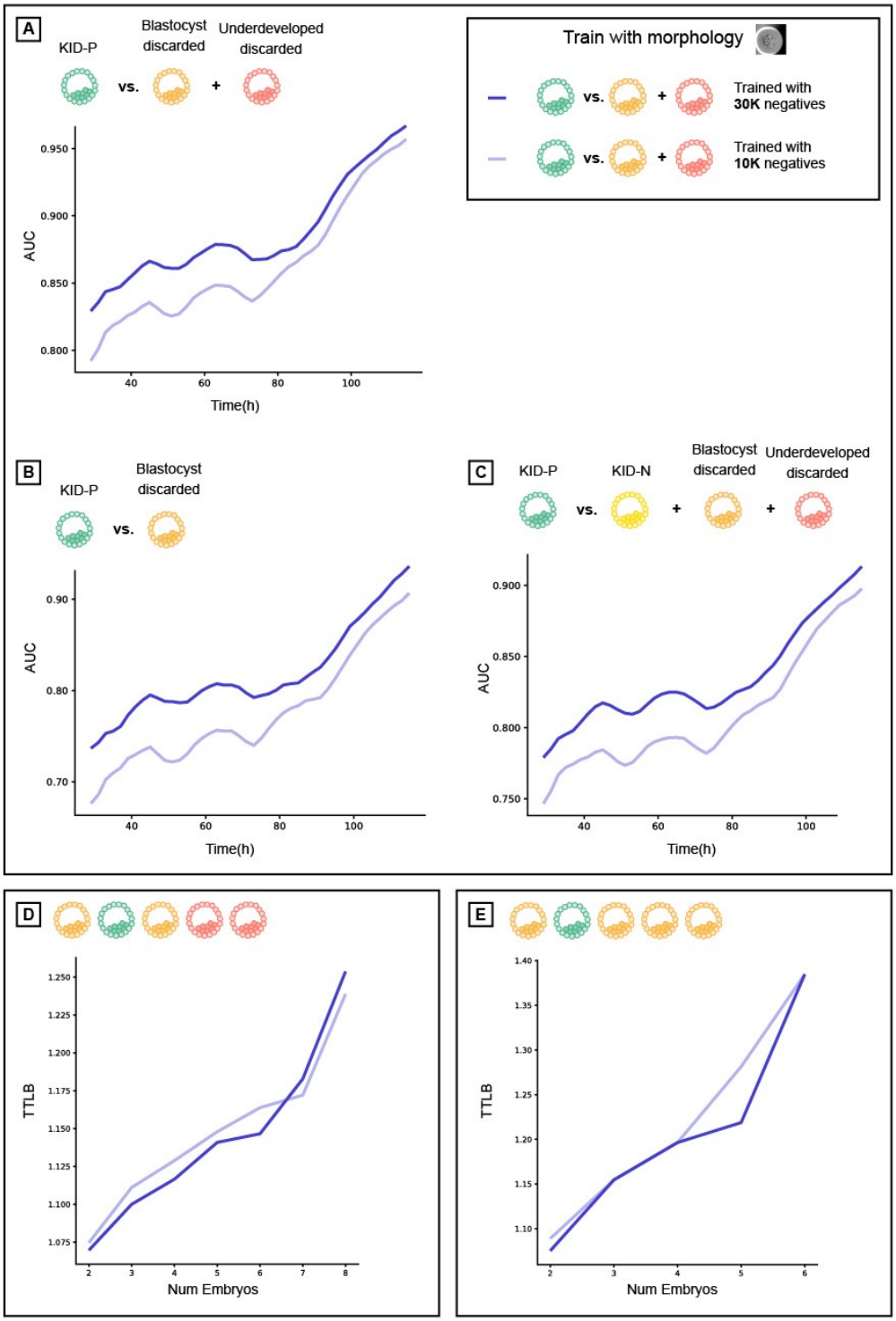


Figure S5: Additional training data enhances morphology-based models. Prediction accuracy: AUC vs. time since fertilization (hours) (**A-C**), average time to live birth (TTLB) (**D-E**) for morphology- based models, trained with different amounts of discarded embryos (negative labels): 10K and 30K, tested over different test sets. Training with more discarded embryos lead to better AUCs for all times and over all test sets, but did not change the TTLB. (**A**) KID-P vs. discarded. (**B**) KID-P vs. blastocyst discarded. (**C**) KID-P vs. KID-N and discarded. (**D**) Ranking KID-P and all discarded. (**E**) Ranking KID-P and blastocyst discarded.

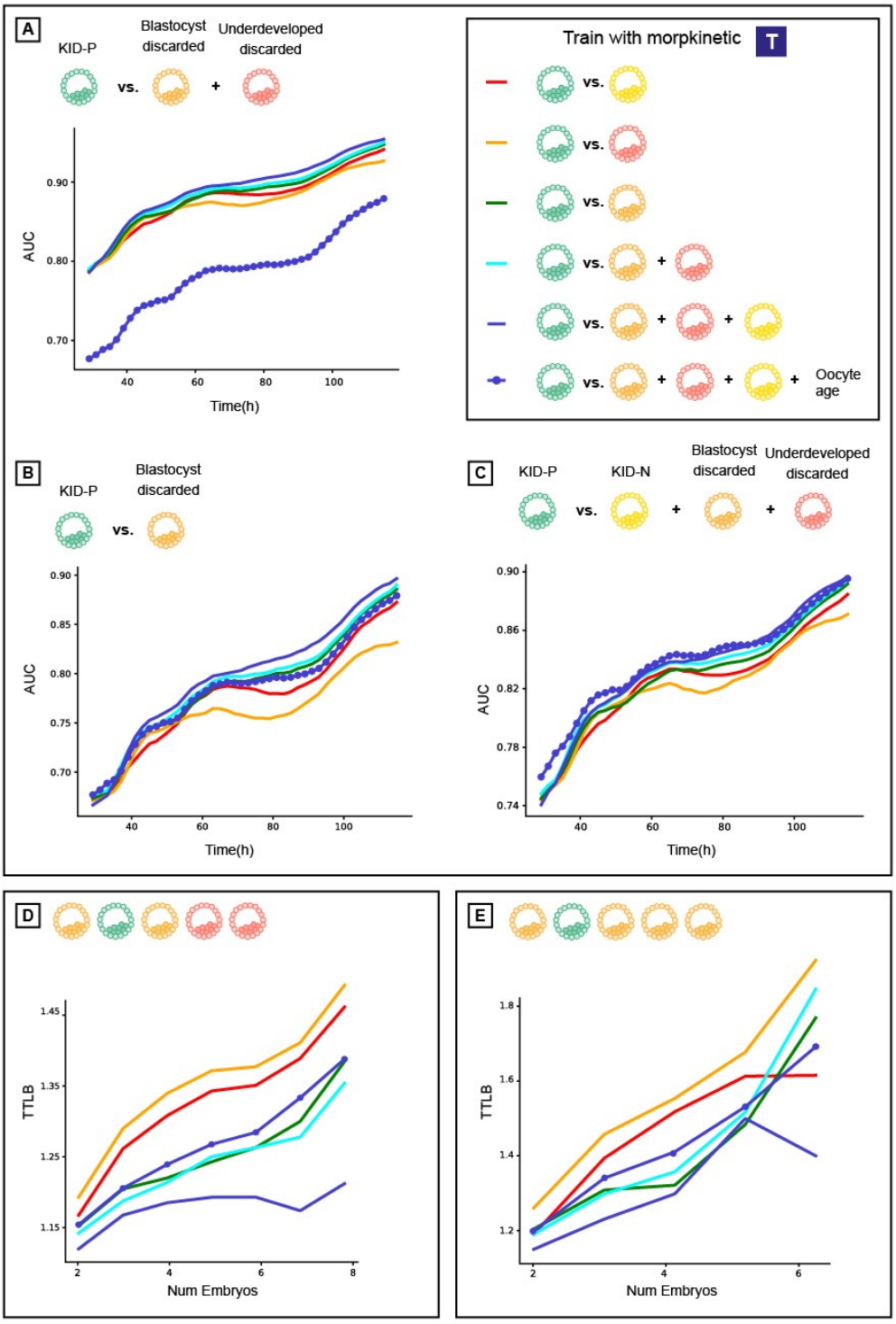


Figure S6: KID-N embryos are not the ideal negative label for solving the task of embryo ranking. Prediction accuracy: AUC vs. time since fertilization (hours) (**A-C**), average time to live birth (TTLB) (**D-E**) for morphokinetic- based models, trained over different negatives: KID-N, discarded, blastocyst discarded, underdeveloped discarded, and KID-N+discarded, tested over different test sets. (**A**) KID-P vs. all discarded. (**B**) KID-P. vs. blastocyst discarded. (**C**) KID-P vs. KID-N and all discarded. (**D**) Ranking KID-P and discarded. (**E**) Ranking KID-P and blastocyst discarded.

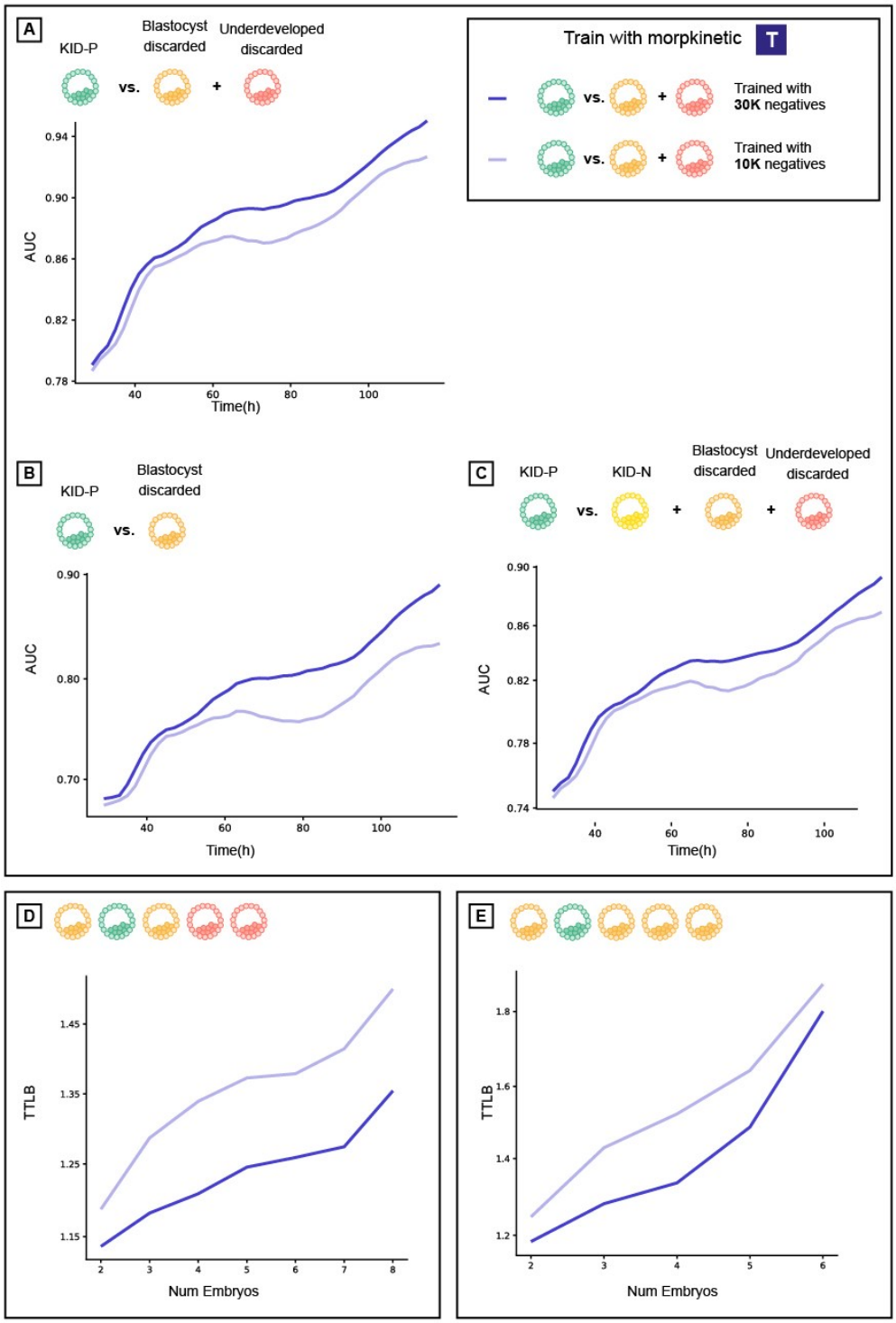


Figure S7: Additional training data enhances morphokinetic-based models. Prediction accuracy: AUC vs. time since fertilization (hours) (**A-C**), average time to live birth (TTLB) (**D-E**) for morphokinetic- based models, trained with different discarded amounts: 10K and 30K, tested over different test sets. Training with more negatives lead to better AUCs and yields to shorter TTLB over all test sets and times. (**A**) KID-P vs. discarded. (**B**) KID-P. vs. blastocyst discarded. (**C**) KID-P vs. KID-N and discarded. (**D**) Ranking KID-P and discarded. (**E**) Ranking KID-P and blastocyst discarded.

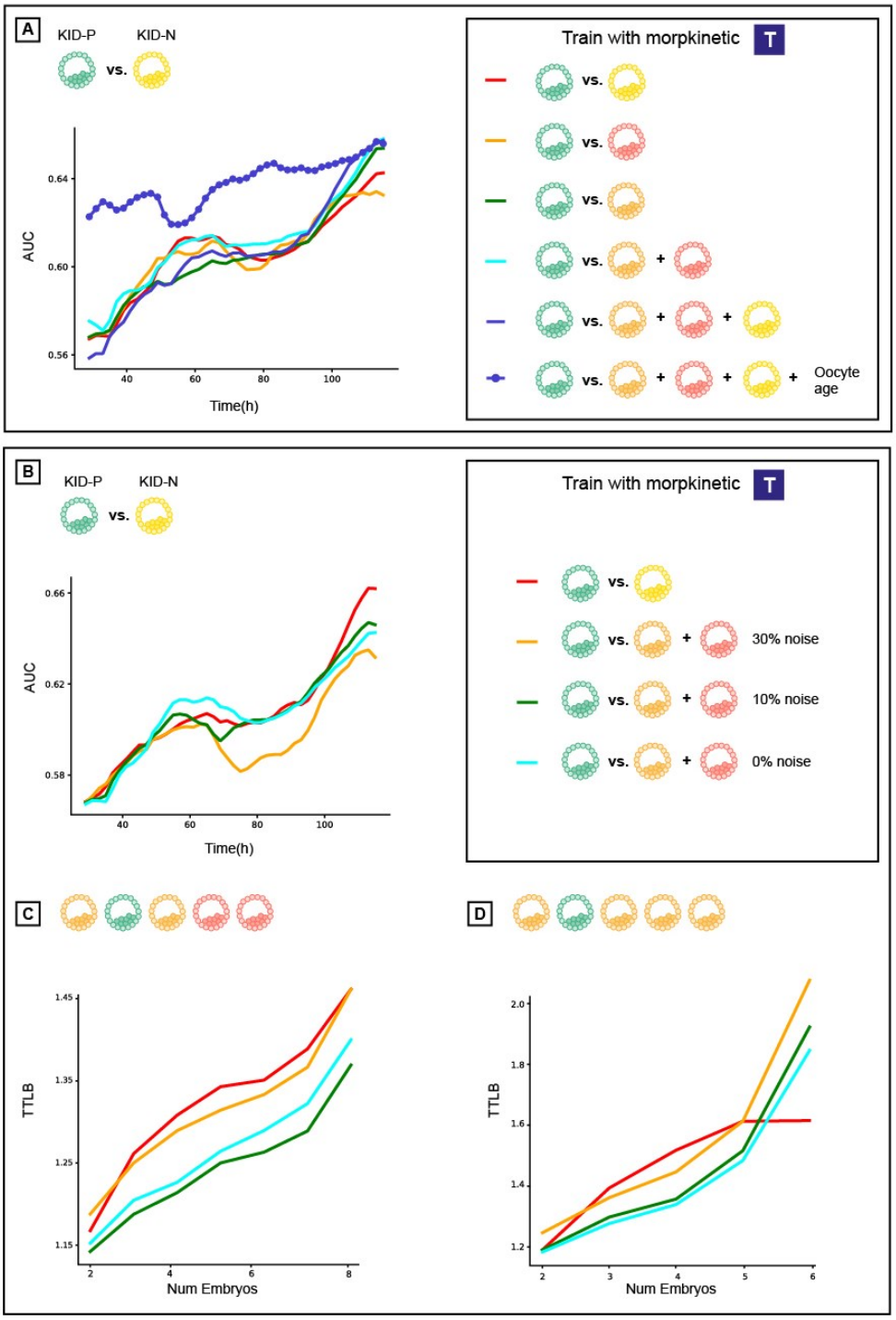


Figure S8: An analysis of optimizing implantation prediction accuracy by trading off label noise level and training data quality based on morphokinetic features. (**A**) Comparison of models trained on different negatives: KID-N, discarded, blastocyst discarded, underdeveloped discarded, and KID-N+discarded (with and without oocyte age) over KID-P vs. KID-N test set. (**B**) AUCs for models trained with known noise levels (of positive labels) over KID-P vs. KID-N test set (**C**) TTLB for models trained with known noise levels (of positive labels) over KID-P vs. discarded test set (**D**) TTLB for models trained with known noise levels (of positives labels) over KID-P vs. blastocyst discarded test set.

Supplementary tables legends

	#cycles	KID-P	KID-N	Blastocyst discarded	Underdeveloped discarded
Train/Val	7102	1314	6485	6603	29904
Test	802	380	637	679	2289

Table S1. Data table. Divided into Train/Validation and Test sets, where all embryos from a given cycle where either used for train, validation or test.

	Positive labels	Negative labels available	Negatives labels for training
KID-P vs. KID-N	1314	6485	6485
KID-P vs. Underdeveloped discarded + Blastocyst discarded	1314	36507	36507
KID-P vs. Blastocyst discarded	1314	6603	6500
KID-P vs. KID-N + Underdeveloped discarded + Blastocyst discarded	1314	42992	42992
KIDp vs. Underdeveloped discarded	1314	29904	6500
KID-P vs. KID-N Learned with oocyte age	1314	6485	6485

Table S2. Train splits for the different trained models. The same amount of negative labels from each embryo discarded category was maintained across all sets of training in order to compare training with different negatives fairly. KID-N consists of the smallest amount of 6500 embryos, so 6500 embryos were randomly selected from all other negative sets (Underdeveloped discarded + Blastocyst discarded, Blastocyst discarded, KID-N + Underdeveloped discarded + Blastocyst discarded and Underdeveloped discarded), first considering all valid negatives, and then selecting 6500 embryos at random from the valid set.

	Positive labels	Negative labels
KID-P vs. KID-N	380	637

KID-P vs. Underdeveloped discarded + Blastocyst discarded	380	2968
KID-P vs. Blastocyst discarded	380	679
KID-P vs. KID-N + Underdeveloped discarded + Blastocyst discarded	380	3605
	Positive labels	**Negative labels**

Table S3. Test splits for the different negatives sets.

Supplementary videos legends

Video S1. Temporal evolution of a KID-P embryo. Bounding box marks the automatically detected embryo contour. Time (in hours) is displayed at the bottom-left. The predicted morphokinetic state and morphological-based implantation prediction score (KID-Score, from time = 30 hours) are displayed at the top-left. The KID-P score at the final frame is 1.

Video S2. Temporal evolution of a blastocyst discarded embryo. Bounding box marks the automatically detected embryo contour. Time (in hours) is displayed at the bottom-left. The predicted morphokinetic state and morphological-based implantation prediction score (KID-Score, from time = 30 hours) are displayed at the top-left. The embryo was discarded after 126 hours, reaching the blastocyst (tB) stage, with a KID-P score of -0.95.

Video S3. Temporal evolution of an underdeveloped discarded embryo. Bounding box marks the automatically detected embryo contour. Time (in hours) is displayed at the bottom-left. The predicted morphokinetic state and morphological-based implantation prediction score (KID-Score, from time = 30 hours) are displayed at the top-left. The embryo was discarded after 127 hours, reaching the start of blastulation (tSB) stage with a KID-P score of -1.31.

Video S4. Temporal evolution of a very low-quality underdeveloped discarded embryo. Bounding box marks the automatically detected embryo contour. Time (in hours) is displayed at the bottom-left. The predicted morphokinetic state and morphological-based implantation prediction score (KID-Score, from time = 30 hours) are displayed at the top-left. The embryo stopped developing after reaching the Pronuclei fading stage (tPNf), with a KID-P score of -9.

Pseudo Contrastive Labeling for Predicting IVF Embryo Developmental Potential

I. Erlich*[1,2], A. Ben-Meir[3,2], I. Har-vardi[4,2], J. Grifo[5], F. Wang[5], C. Mccaffrey[5], D. McCulloh[5], Y. Or[6], and L. Wolf[7]

[1]The Alexender Grass Center for Bioengineering, School of computer Science and Engineering, Hebrew University of Jerusalem, Israel
[2]Fairtilty Ltd., Tel Aviv, Israel
[3]IVF unit, Department of Obstetrics and Gynecology, Hadassah Medical Organization and Faculty of Medicine, Hebrew University of Jerusalem, Israel
[4]Fertility and IVF Unit, Department of Obstetrics and Gynecology, Soroka University Medical Center and the Faculty of Health Sciences Ben-Gurion University of the Negev, Beer-Sheva, Israel
[5]New York University Langone Prelude Fertility Center, New York, New York
[6]Fertility and IVF Unit, Obstetrics and Gynecology Division, Kaplan medical center
[7]The School of Computer Science, Tel Aviv University

Abstract

In vitro fertilization (IVF) is typically associated with high failure rates per transfer, leading to an acute need for the identification of embryos with high developmental potential. The current methods are tailored to specific times after fertilization, often require expert inspection, and have low predictive power. Automatic methods are challenged by ambiguous labels, clinical heterogeneity, and the inability to utilize multiple developmental points. In this work, we propose a novel method that trains a classifier conditioned on the time since fertilization. This classifier is then integrated over time and its output is used to assign soft labels to pairs of samples. The classifier obtained by training on these soft labels presents a significant improvement in accuracy, even as early as 30 hours post-fertilization. By integrating the classification scores, the predictive power

*Corresponding author, submitted to *scientific reports*

is further improved. Our results are superior to previously reported methods, including the commercial KIDScore-D3 system, and a group of eight senior professionals, in classifying between multiple groups of favorable embryos to groups that are classified as less favorable based on implantation outcomes, expert decisions based on developmental trajectories, or genetic tests.

1 Introduction

In vitro fertilization (IVF) is the most effective form of assisted reproductive technology. However, IVF treatments are inefficient, costly, lengthy, and place an emotional burden on the future parents. In particular, selecting the embryos with the best implantation potential remains a challenging task. Despite decades of clinical practice, there is no reliable non-invasive method to identify the small fraction of embryos that possess the highest potential to develop into a blastocyst, which can then be implanted and hopefully proceed to term [11]. Thus, to maintain reasonable pregnancy rates, the current practice often involves the transfer of multiple candidate embryos into the uterus. This results in clinical complications and health risks to both the newborn and the mother, with over 10% of IVF pregnancies being twins or more [29].

Several methods have been proposed to rate embryos by their implantation likelihood. The current practice involves a single image of the embryo taken by a microscope mounted camera moments before transferring, typically at day 5, with relatively low identification rates [42]. Methods for earlier development stages are rarely studied, perhaps because it is considered almost impossible to identify viable embryos based on a single pre-blastocyst image [12, 7].

Time-lapse incubation (TLI) has the potential for improving identification due to the availability of more data points and the ability to model embryo development. Time lapse videos typically consist of 70 images/day, 7 foals each. That is, TLI provides approximately 2500 images for each day 5 blastocyst embryo. To date, no accepted or standardized algorithm exploits the added information in the TLI sequence or that can outperform traditional, manual microscopy-based scoring. As a result, despite the added information TLI provides, IVF pregnancy rates remain low.

TLI studies are often limited to later developmental stages, mainly day 5 blastocysts, and require manual annotations of many spatial-temporal features from the TLI stream to obtain a single score [5]. As far as we can ascertain, our method is the only method that can be applied to all embryos at any time or developmental stage.

Our work integrates classification information over time to benefit from the availability of multiple images in TLI during training, and can be applied to either TLI data or single microscopy images during test. We do not emphasize tracking and meeting explicit development milestones. Instead, we focus on obtaining a stable classification score by building shared feature layers for the multiple development stages and making use of unlabeled or ambiguously

labeled embryos. The latter is crucial, since many embryos were not transferred, and even those that were, are typically transferred as part of a group of embryos, for which the fate of each individual embryo is uncertain.

Our semi-supervised method is based on assigning pseudo-ranking labels to pairs of embryos. A classifier is then trained based on these labels, by employing a Siamese network. Information is further integrated by considering multiple time points. The improved labels are then used to obtain a second set of pseudo labels and the process repeats one more time.

Our results indicate that temporal integration helps, even if it is employed only during training, that employing pseudo labels together with pair contrasting is effective, and that the obtained method greatly outperforms the current commercial system.

2 Related Work

In the early days of IVF treatments, a two dimensional snapshot was taken by a microscope at the different development stages of the embryo [12]. Many machine based models have been proposed involving clinical features, such as oocyte age, causes of infertility, oocyte stimulation, and semen analysis, in addition to embryonic data [44, 32].

In particular, several (typically 10-100) features are chosen, which are then subjected to a feature selection technique such as SVM [15], multivariate logistic regression [8], genetic algorithm and/or decision trees [14], Bayesian networks [10] and even ANN [24]. Nevertheless, all these methods mainly concentrate on predicting the probability of a patient to conceive, rather than ranking which embryo has the best probability of implanting given a group of possible embryos and their corresponding images. Several models have been proposed for ranking embryos based on a single image [44, 32], taken at a specific time such as day 3, cleavage-stage [18], day 5, blastocyst stage [15] or even day 1, when the pronuclei appear [28].

To overcome the inability to accurately evaluate early stage embryos, culturing embryos to the blastocyst stage combined with grading protocols have frequently been implemented in both research and clinical practice [12, 34, 40, 30]. Several AI methods were recently presented. STORK [19] automates and ERICA [6] replaces the manual ranking of late blastocyst embryos. For developing STORK , expert embryologists were asked to grade embryos using the Veeck and Zaninovich system [45], a slightly modified version of the Gardner system [12], as poor, fair or high quality. Then, a CNN was trained to distinguish between the high quality and the poor quality embryos, disregarding the fair quality embryos. They reported an AUC of 0.98 predicting whether an embryo was of good quality or poor quality over a test set tagged by one of the Cornell clinic's embryologists. However when the embryos were tested by embryologists in a different hospital (Universidad de Valencia, Valencia) the AUC dropped to 0.75, suggesting that STORK overfitted to the blastocyst grading at the Cornell clinic. Moreover, fair quality [20] accounts for more than 40% of the STORK

data and is the hardest to identify. Finally, the STORK system implements a scoring system rather than attempts to predict implantation outcomes. Recent studies suggest that the Gardner score is not well aligned with implantation probability or even euploidy [6]. ERICA [6] is a two-step method with the goal of ranking late stage blastocysts. In the first step, 96 spatial features are extracted from a 2D image. This is followed by a second step, in which a CNN is trained over these 96 features to predict embryo ploidy and implantation. We note that the spatial features are not exactly defined and require human expertise. ERICA was reported to obtain an accuracy of 70%. However, the train and test embryos were hand-picked as having "good" images; i.e. embryos of lower quality were ignored. Furthermore, the sample only consisted of 84 embryos from 19 patients and the test set was biased by a criterion of having at least one euploidy and one aneuploidy embryo. This may imply a skew towards younger patients [9, 10, 14] since patient age was shown to predict implantation [9]. In addition, some of the embryos were tagged as positive based solely on beta human chorionic gonadotropin (beta-hCG) results, which is not sufficient, since implantation is normally defined as the presence of (at least) a gestational sac or a heartbeat [19, 32], to exclude early miscarriages (chemical pregnancy). Most importantly, more than 95% of the embryos in their study were either in the middle or after the hatching stage, which typically corresponds to Day 6 or Day 7 embryos, while the vast majority of today's transfers happen earlier.

Time-lapse incubation (TLI) enables the continuous tracking and screening of fertilized embryos, unlike traditional microscopy that is limited to snapshots of a few discrete points in time [3, 26]. Time-lapse imaging overcomes the traditional drawbacks of traditional microscopy, such as exposing the embryos to environmental changes [25]. It is known that time-lapse microscopy can point to new dynamic markers of embryo quality and provide novel algorithms for effective embryo selection [4].

Several algorithms were proposed in the last few years based on embryo kinetics and morphology as identified manually by experts and have been used to predict implantation [17, 26, 27], blastocyst formation [27] and even genetic chromosomal disorders [33, 41]. However, time lapse machines have yet to yield better prediction results than regular microscopy [4]. Moreover, current studies are based on morphology and morphokinetic parameters which require manual annotation and are thus associated with intrinsic variability.

A few studies have addressed the task of blastocyst formation but have mainly considered embryos that were observed for 5 days, by which time 97% of the blastocyst embryos had already developed into blastocysts [24]. These authors reached an AUC of 80% with the blastocyst scoring algorithm, which however was found to be inferior to manual accuracy by visual inspection. Several studies have reported positive correlations between pronuclei markers and embryo viability. Specifically, markers such as the appearance of pronuclei and fading timing (tPNa, tPNf), the number of pronuclei, pronuclei shape, symmetry, and joint path were found to be associated with blastocyst development and implantation outcomes [1, 38, 21]. However, these markers have never been integrated into a scoring system, and not all of the embryos in these studies

were placed in the TLI early enough to track the pronuclei stage.

Other studies have addressed the task of cleavage stage (day 3) classification using time lapse imaging [17, 26]. However, these do not consider events after 76 hours. A study by Liu et al. [23] presents a flow diagram with six conditions for scoring the embryo with a grade ranging from A to F after 68 hours.

The most common (patented) implantation scoring for cleavage stage embryos is EmbryoScope's built-in algorithm (Vitrolife) called KIDScore-D3, which gives a (one-time) prediction after 66 hours. KIDScore-D3 implements a binary decision tree that is based on five morphokinetic conditions. The authors report an AUC of 0.65 in predicting implantation over day 3 transfers [31]. Another built-in score, called KIDScore-D5 combines both morphokinetic with morphological assessment and provides a score between 1 and 6. A recent study [13] suggested that the KIDScore-D5 improves implantation rates, compared to traditional morphological scoring, and is associated with PGT euploidy. It was not released to the public, so we cannot evaluate its performance. Crucially, all these studies are tailored for specific timing or were based on manually selected features that are only present in a small fraction of the TLI stream making them hard to use in practice, given the high workload and required expertise.

Other recent studies have suggested new markers and grading models for day 5 blastocyst embryos based on TLI tracking. One study reported a strong correlation between spontaneous blastocyst collapse and pregnancy outcomes [36]. In another study, it was shown that TLI streaming can be exploited to automatically classify Inner Cell Mass (ICM) and Trophectoderm grading [20]. The results indicated that TLI led to superior results compared to a single microscopy image snapshot. Another study employs different spatial parameters manually acquired at different time/ development stages (mainly late blastocyst stages) to accurately predict implantation results [5]. However, there is yet to be found a quantitative scoring algorithm based on these findings.

The first deep network for predicting the implantation potential of blastocyst embryos from time-lapse videos was recently presented [43]. Unfortunately, the way the entire video sequence was fed to the network was not described in this study. The authors reported an AUC of 93%. However, considering the distribution of the dataset (694/1079/7063 KID-positive / KID-negative /Discarded), the results correspond to a model that accurately discarded 100% of the discarded embryos and predicted almost randomly (with a 55% success rate) over all the remaining embryos (i.e., embryos with KIDp/KIDn label). Since by definition all the discarded embryos could have been discarded manually, these results reflect a 55% accuracy for the KIDp/KIDn embryos. Further, it could only be evaluated over blastocyst embryos, so this algorithm cannot be used as a grading system as of the early stages of embryonic development (Days 0-3).

3 Method

Each image (500x500 pixels) depicts a centered spheroid culture well of a size that is approximately 430x430 pixels. The embryo itself can be found anywhere

inside the well. Thus, a segmentation of the embryo sub image is used as a pre-process, see Fig. 1. This segmentation employs a UNet [35] network trained to minimize a pixelwise soft hinge loss function.

Since the embryo consists of high variability due to temporal changes, from one cell to fully hatching, it is required to segment exactly which pixels belong to the embryo. The imbalance between the "foreground" pixels (i.e, that contain the embryo) and "background" pixels, and the varying difficulty between pixels of each type, calls for a careful weighting of the training objective. Inspired by the focal loss [22], our objective function multiplies two terms per pixel. The first is a binary soft hinge loss. It is obtained by performing the soft-max operation over the hinge-loss, which is both smooth and upper bounds the 0-1 loss [39].

$$l_p(s_p, y_p) = \log\left(1 + e^{\gamma(m - y_p \cdot s_p)}\right) \tag{1}$$

where y_p is the mask grand-truth at the pixel p, which denotes if the pixel is inside the embryo or not, and s_p is the prediction score at the pixel p. The margin m and the constant γ are both set to 1.

The second term weights all the pixels, such that pixels that are solved with high confidence are underweight. To compensate for the imbalance between foreground and background, each type is separately weighted to have 1 unit weight.

$$w_p(s_p, y_p) = \frac{e^{-s_p \cdot y_p}}{\sum_{q \in [P], y_q = y_p} e^{-s_p \cdot y_p}} \tag{2}$$

The overall loss per single example is given by,

$$l = \sum_{p \in [P]} w_p\left(s_p, y_p\right) \cdot l_p\left(s_p, y_p\right) \tag{3}$$

The training was performed over 50634 manually tagged images from 455 TLI embryo movies and validated over a test set of 9511 images from 90 TLI embryo movies. The resulted per pixel 0-1 accuracy is 0.994 with the weighting scheme applied, and 0.991 without. Using the AUC measure, the corresponding results are 0.994 vs. 0.984 on the test set. This indicates high accuracy segmentation, even without the modified loss, which is further improved (eliminating a third of the errors) by it.

The IVF data is extremely heterogeneous and the labels are noisy. First, the reasons for infertility and implantation failure vary between cases and depend on both the embryo and the mother. Second, the Known-Implementation-Data (KID) provides partial labeling: the successful implantation cases (KIDp) indicate a healthy embryo, while a negative outcome (KIDn) may indicate issues that may not be associated with the embryo itself. Third, the data consists of embryos captured at multiple time points, at completely different levels of development. Images taken at day 2 typically consist of 4-8 cells, while images taken at day 5 are typically in some of the blastocyst stages.

To tackle the first issue, we use the age of the mother (more precisely the oocyte age) as an instrumental variable that is correlated with the underlying medical challenge and compare embryos for mothers of similar ages.

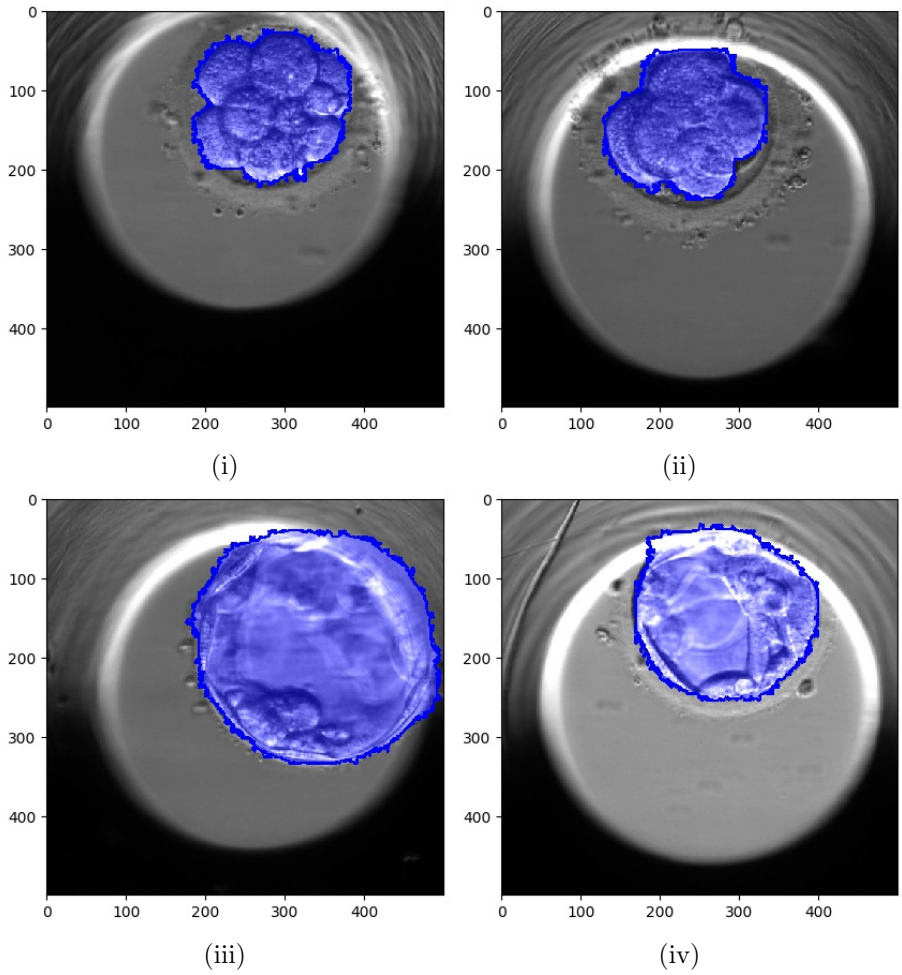


Figure 1: Embryo pixelwise segmentation using a UNet [35]. The output masks the network outputs are given for different developmental stage examples. (i) 8 cells, (ii) 10 cells (iii) blastocyst (iv) expanded blastocyst.

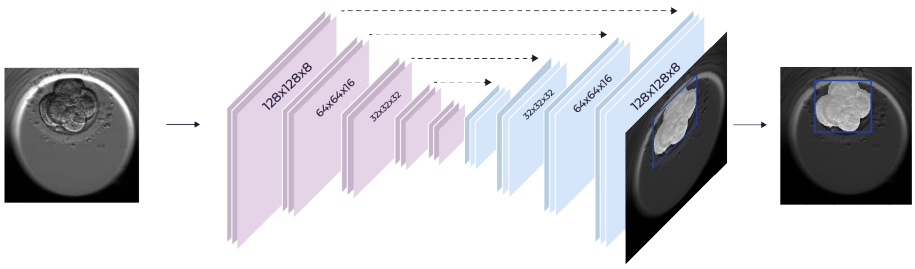


Figure 2: The classification model. A 256x256 cropped embryo input image, is fed into Resnet50 CNN with multiple prediction heads. Each head corresponds to a temporal window of 2 hours.

To tackle the second issue, we provide a method that relies on soft labels. This also allows us to learn from embryos for which the KID status is unknown. The soft labeling is applied to all pairs of samples $i, j \in [N]$ in a minibatch of size N, that contains embryos of mothers of similar age.

$$L = \sum_{i \neq j \in [N]} l\left(s_{i,j}, y_{i,j}\right) , \tag{4}$$

where $s_{i,j}$ is the joint score of the pair (i, j) in mini-batch at size N. $y_{i,j}$ is the label associated with the pair, and l is some loss function.

To tackle the third issue, we condition the prediction on the stage of development by employing multiple prediction heads in our network classifier, where each head corresponds to a temporal window of two hours within the time interval $\Delta_t = [t_0, t_e]$ we inspect.

Our network models have a ResNet50 [16] backbone, after which a separate classification-head of two residual convolutions is dedicated to each time window, see Fig 2. At inference time, we only consider the head that is relevant to the frame we classify, thus conditioning our result on the time of capture.

To take advantage of the fact that we have multiple images per embryo, we integrate information across different time points, making our soft labels more informative. Specifically, we consider the time interval $\Delta_t = [t_0, t_e]$ and integrate the classifier scores over time using an auto-regression moving average (ARMA) model, in order to obtain a fused score.

With these building blocks, we employ six models A–F, defined sequentially, on top of each other, see Fig 3.

Model A is a binary classifier A that accepts single image x^t from all time points $t \in \Delta_t$. The train data is a set of the form $\{(\{x_i^t\}_t, y_i)_i\}$ where y_i is the KID label of the i-th embryo. The loss function is a the same soft hinge loss given in Eq. 1.

$$L_h = \sum_{i \in [N], t \in \Delta_t} \log\left(1 + e^{\gamma\left(m - y_i A(x_i^t)\right)}\right) , \tag{5}$$

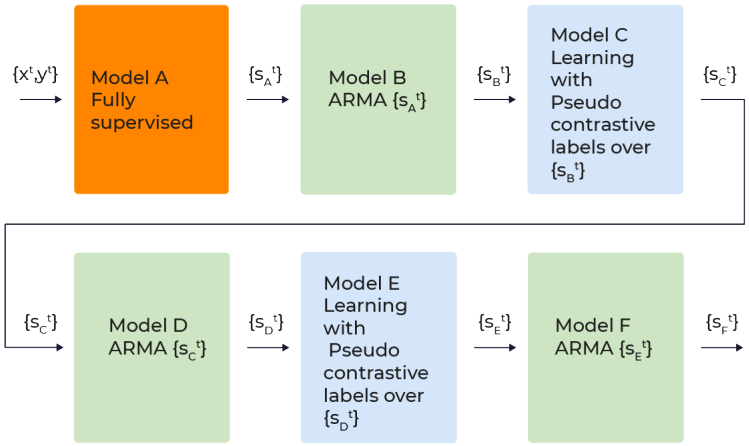


Figure 3: Models overview. (A) A CNN model, trained in a fully supervised manner over KIDp vs. KIDn, a total of 7799 embryos. (B) A temporal ARMA filtering over the outcome of A. (C) A CNN model, trained using pseudo constrictive labels obtained by the outcome of B, over all training embryos, a total of 61581 embryos (D) An ARMA filtering over the outcome of C. (E) A second iteration of pseudo constrictive labels: a CNN model, trained using pseudo constrictive labels obtained by the outcome of D. (F) An ARMA filtering over the outcome of E.

where m and γ are parameters that control the margin between positives and negatives and loss softness, respectively.

Model B integrates the single frame outputs $A(x_i^t)$ using a first order ARMA model. Namely, for all $t \in \Delta_t$, $B(x_i^t) = \alpha A(x_i^t) + (1-\alpha)A(x_i^{t-1})$. Embryo score $B(x_i^{t_e})$ is given by considering the last time point.

Model C is a neural network C that accepts an image as input. It is trained using all embryos, including those for which KID data is not available (embryos that were not transferred or for which the outcome is ambiguous).

Model B provides the pair metric $s_{i,j}$ and pair label $y_{i,j}$ from Eq. 4. The former is given for a pair of input images $x_i^{t_i}, x_j^{t_j}$ as

$$s_{i,j} = C(x_i^{t_i}) - C(x_j^{t_j}), \tag{6}$$

$$y_{i,j} = \begin{cases} 1, & B(x_i^{t_e}) - B(x_j^{t_e}) > \theta, y_j \neq 1 \\ -1, & B(x_j^{t_e}) - B(x_i^{t_e}) > \theta, y_i \neq 1 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The CNN C is, therefore, trained based on the multi-frame scores of model B, which is applied to samples with KID positive (KIDp), KID negative (KIDn), KID unknown (KIDu), or samples with no KID tag (not-KID). These soft labels are restricted such that KIDp samples would not be on the negative side of the soft label. Note the pairs are not limited to having the same time, which encourages early classification. The threshold θ is meant to relax errors in model B. Overly large θ may dismiss too many pairs. On the other side, a low θ may overfit to the errors of model B.

To avoid the implantation bias that is associated with oocyte age, out of all pairs in a given minibatch, we only consider pairs that have an oocyte age that is smaller or equal to a threshold θ_o.

Model D is to model C what model B is to model A. It integrates C's outcome over the interval Δ_t, $D(x_i^t) = \alpha C(x_i^t) + (1-\alpha)C(x_i^{t-1})$. **Model E** applies the soft label procedure of Eq. 4 to model D, the same way that model C employs the outcome of model B. Finally, **Model F** integrates model E using the ARMA model, similar to model D with respect to C and model B with respect to A.

4 Experiments

Data After the culturing period inside the incubator concludes, the best embryos (typically 1-2, out of 10-20) are transferred. The remaining embryos are frozen, if they appear morphologically viable, or otherwise discarded. For those embryos that were transferred, tagging follows the Known Implantation Data (KID) designations: (i) KID-positive (KIDp), if the number of transferred embryos equals the number of gestational sacs, i.e., each transferred embryo was successfully implanted. (ii) KID-negative (KIDn), if there are no gestational sacs, i.e., none of the transferred embryos implanted. (iii) KID-unknown (KIDu), if the number of gestational sacs is greater than zero but smaller than

Table 1: Train/validation (val) and Test embryo implantation data.

	KIDp	KIDn	KIDu	Discarded	Hard Discarded	not-KID	Euploid	Aneuploid
Train/val	1314	6485	2106	36507	11504	14934	342	713
Test	380	637	0	2289	687	0	59	117

the number of transferred embryos. In this case, the outcome of each individual embryo is uncertain.

Untransferred Embryos (frozen or discarded) have no tag and are denoted not-KID. Typically, those account for ∼85% of the embryos [43]. We denote embryos as *hard-discarded*, discarded that developed at least to the blastocyst stage. Embryos that had preimplantation genetic testing (PGT) have in addition a tag of *euploid*/*aneuploid* for chromosomal normal/abnormal, respectively.

The training data was taken from four hospital centers (left out for anonymity). After fertilization, the embryos were incubated in EmbryoScopeTM (Vitrolife, Copenhagen, Denmark), a time-lapse incubator that captures images every 15 to 20 minutes. The embryo development videos were collected from July 2014 to December 2019, with oocyte ages of 23.6-43.6 years. Another set used, consists of embryos that underwent PGT from another anonymous center. The study was approved by the Investigation Review Board of the anonymous institute. Data was spit to Train and Test, without intersection between patients, such that the test set is composed only of embryos for which data is available for the entire length of 114 hours. This makes the KIDn test set harder (since easier negative cases from a day 3 transfer would have been discarded and not transferred on day 5) and it allows for comparing results between different time points on the exact same test set. The sizes of the different splits are detailed in Table. 1.

Setting the hyperparameters Setting $m = 0$ and $\gamma = 1$ of L_h, the binary soft hinge loss would become the binary cross entropy loss. In our work, we set early on the development process $m = 1, \gamma = 1$, which add a significant margin, and further regularize with weight decay of 1e-5. The batch size was set to 24, thus allowing a maximum of 264 pairs in a single minibatch. The average oocyte age of KIDp/KIDn embryos was 31.33/35.78 with Standard Deviation (SD) of 4.98/5.97, respectively. We employ $\theta_o = 2$, avoiding comparisons between oocytes that have more than a half SD age gap, and having an average of 32 valid pairs in a minibatch.

The parameter θ in models C and E was set by considering the two 100-bin histograms computed for the values obtained by applying model B to the training sets of KIDp and KIDn. Specifically, the value is the mean over the difference between the mean value of matching bins of the two histograms. A 2nd value of θ is used in learning model E based on Model D, and it is computed by considering the histograms obtained from applying model D on the two training sets.

The ARMA coefficient α used by integration models B,D,F was set to 0.05, to incorporate long scores memory.

Training involved all training data, whereas the results were tested with respect to four different subsets of the test set: (i) KIDp vs. KIDn, (ii) KIDp vs. aneuploid, (iii) KIDp vs. Discarded, and (iv) KIDp vs. Hard Discarded.

Table 2: Classification results (AUC) for each model for the different datasets (each denoted by the negative class) at the end of day 3 and at the end of day 5.

Model	Day	KIDn	Aneuplody	Discarded	Hard-Discarded
A (multi class model)	3	0.561	0.180	0.718	0.590
	5	0.653	0.633	0.916	0.849
B (integration of A)	5	0.656	0.67	0.927	0.868
C (pair model over B)	3	0.620	0.604	0.865	0.780
	5	0.669	0.873	0.965	0.924
D (integration of C)	5	0.671	0.885	0.967	0.929
E (pair model over D)	3	0.624	0.724	0.861	0.780
	5	0.678	0.890	0.967	0.935
F (integration of E)	5	**0.681**	**0.904**	**0.970**	**0.942**
Liu et al. [23]	3	0.594	0.691	0.814	0.685
KIDScore-D3 [31]	3	0.582	0.671	0.832	0.707

We evaluate the accuracy of soft-labeling (models C-F) across all datasets, times, and ages, compared to both fully supervised learning (models A-B) and two previously reported studies, Liu et al. [23] and the FDA-approved KIDScore-D3 [31].

The results are reported in Table. 2. Evidently, each progression of our method (from A to B to C, etc.) improves the classification accuracy, for both day 3 and day 5 and across the four classification tasks. This indicates that integration over time helps (A to B, C to D, E to F) and the utility of training the single image classifiers (C, E) using information integrated over time (B and D, respectively).

Our latest single-image classifier (model E), which does not observe multiple images during test (it is trained based on sequence information), outperforms on day three both Liu et al. and the proprietary day 3 KIDScore-D3 models, despite both latter models requiring 66-68 hours of continuous monitoring and manual annotations.

To better understand the availability of information across different time points, we evaluate models A, C, and E at different times, see Fig. 4. Also shown is the performance obtained by the Liu et al. and KIDScore-D3 models, after observing multiple measurements over three days. As can be seen, model E almost always improves upon model C and always improves upon model A. Both Liu et al. and KIDScore-D3 models are only competitive with our initial model (A), and with our intermediate model (C) when considering KIDp vs. aneuploid (see Supplementary for the two other tasks) before 76 hours since fertilization. Notably, as time progresses, the accuracy tends to increase. Around day 4,

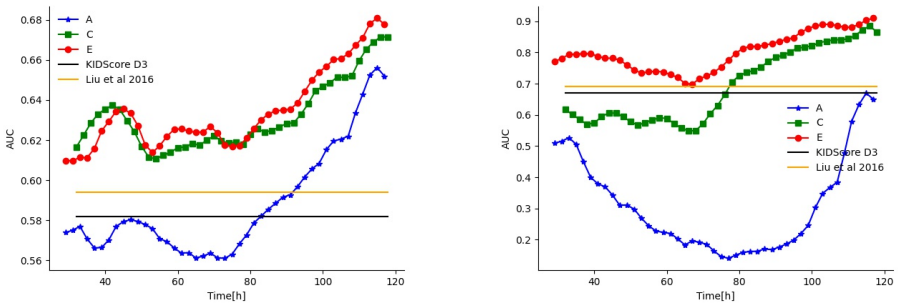


Figure 4: Prediction ability (AUC) as a function of time since fertilization (hours) for single-frame models. (a) KIDp vs. KIDn classification, (b) KIDp vs. aneuploid.

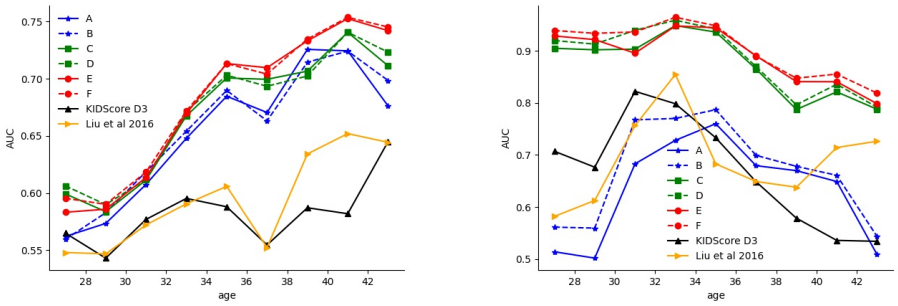


Figure 5: Prediction ability (AUC) as a function of oocyte age (years). (a) KIDp vs. KIDn classification, (b) KIDp vs. aneuploid.

Table 3: KIDp vs. KIDn Classification results (AUC) of multiframe model F compared to professional embryologists at the end of day five, for different ages of the mother. The mean is computed across the nine age groups. The p-value is obtained with the unpaired t-test.

Model/Annotator	age 27	age 29	age 31	age 33	age 35	age 37	age 39	age 41	age 43	Mean
embryologist 1	0.562	0.555	0.627	0.665	0.635	0.580	0.667	0.725	0.734	0.639
embryologist 2	0.613	0.599	0.623	0.638	0.631	0.599	0.661	0.686	0.717	0.641
embryologist 3	0.61	0.612	0.643	0.682	0.656	0.634	0.699	0.703	0.684	0.658
embryologist 4	0.590	0.574	0.591	0.621	0.610	0.602	0.661	0.687	0.697	0.626
embryologist 5	0.564	0.589	0.634	0.687	0.664	0.647	0.677	0.693	0.683	0.649
embryologist 6	0.632	0.607	0.625	0.624	0.599	0.594	0.653	0.683	0.675	0.632
embryologist 7	0.613	0.575	0.589	0.633	0.586	0.569	0.622	0.654	0.693	0.615
embryologist 8	0.632	0.589	0.636	0.666	0.662	0.617	0.657	0.683	0.731	0.652
Mean embryologists	0.602	0.587	0.621	0.652	0.630	0.605	0.662	0.689	0.702	0.639
SD embryologists	0.026	0.018	0.019	0.024	0.028	0.025	0.020	0.019	0.021	0.013
model F	**0.636**	**0.638**	**0.663**	**0.703**	**0.711**	**0.687**	**0.738**	**0.760**	**0.755**	**0.699**
P-value	0.010	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000

which is typically corresponding to the morula stage, there is often a small drop in performance. On the contrary, starting the blastulation phase, the accuracy rate significantly increases. Both trends are well-known in the literature [2, 37].

The age of the oocyte is an indicator of the underlying fertility issue. In Fig. 5, we separate the results based on this age. As can be seen, all of our models outperform both Liu et al. and KIDScore-D3 models when predicting the KID status across all age groups, and the models that involve pair-training (C-F) outperform it in all groups also for aneuploid detection.

See supplementary for an ablation study that compares pair loss with binary cross entropy loss.

A comparison to human graders The performance of multiframe model F was further compared to eight professional embryologists, each from a different clinic, spread around multiple centers globally.

Each embryologist was asked to score the embryos that were transferred, e.g. KIDp vs. KIDn, between 1 to 5 (higher is better). Then, an AUC was calculated with respect to the implantation tag, for each age group. The results comparing our last model to the embryologists are reported in Table. 3. As can be seen, the model outperforms the human embryologists across all age groups at a p-value that is 0.01 at most, often much lower.

5 Discussion

One of the most challenging barriers in solving the task of embryo identification is the lack of labeled data. Typically, only 15% of the embryos are transferred.

The others are frozen or discarded. Multiple embryos are often jointly transferred, leading to ambiguous labeling. Moreover, embryos may not implant successfully due to factors that are not embryo-related. This, in turn, means that negative labels are noisy. We tackle both problems by introducing a pseudo soft label scheme, which requires only part of the data to be labeled, for initializing the learning process.

The resulting single-frame classifier learns a time-dependent prediction that can be applied to all embryos with no limitation of time or developmental stage. Further, it does not require any manual annotations which are time consuming and provides continuous estimations of embryo viability as early as day 2 (Fig. 4), rather than a single score long after the oocyte was fertilized. Integration of the single frame scores is shown to further improve the results. Thus, unlike other algorithms, it can be used seamlessly in real-time IVF workflows, using even low-cost conventional microscopes. Further, it was shown to outperform both Liu et al. [23] method and Vitrolife TLI's built-in scoring, the KIDScore-D3, in identifying KIDp embryos vs. KIDn, aneuploid, and hard-discarded embryos, at all times, starting at 30 hours, which is 36 hours before KIDScore-D3 and 38 hours before Liu et al. are available. A more detailed comparison yields that this superiority is kept for all ages and datasets (Fig. 5 and supplementary) for models C onward.

The presented single-image classifier successfully identified KIDp vs. aneuploid with an AUC of 0.89, which implies that it can become a non-invasive PGT replacement. We note that the classification of aneuploid vs. KIDp is more clinically relevant than the outcome of the PGT test itself, since not all the euploid embryos can produce implantation. Additional discussion of the results can be found in the supplementary.

6 Conclusions

The literature on fully automatic predictive IVF does not address the utilization of multiple frames taken over time and does not provide solutions for utilizing, during training, embryos with missing or ambiguous labels. We address the first problem by proposing a simple time integration method. The obtained knowledge is further used to train a single-frame classifier that can be used, even if multiple frames are not available. To learn in a semi-supervised manner that incorporates unlabeled and ambiguously labeled samples, and to address the problem of negative implantation labels for high-potential embryos, we suggest a pseudo-labeling scheme that is applied to pairs of samples. Our results indicate that both the time integration and the pseudo labeling improve results, even when applied multiple times, in an interleaving manner. Our method obtained considerably better performance in comparison to the existing FDA-approved system, which requires expert labeling across multiple time frames.

References

[1] Jesús Aguilar, Yamileth Motato, María José Escribá, María Ojeda, Elkin Muñoz, and Marcos Meseguer. The human first cell cycle: impact on implantation. *Reproductive biomedicine online*, 28(4):475–484, 2014.

[2] Mina Alikani, Gloria Calderon, Giles Tomkin, John Garrisi, Magdalena Kokot, and Jacques Cohen. Cleavage anomalies in early human embryos and survival after prolonged culture in-vitro. *Human reproduction*, 15(12):2634–2643, 2000.

[3] Belén Aparicio-Ruiz, Laura Romany, and Marcos Meseguer. Selection of preimplantation embryos using time-lapse microscopy in in vitro fertilization: State of the technology and future directions. *Birth defects research*, 110(8):648–653, 2018.

[4] Natalia Basile, Mauro Caiazzo, and Marcos Meseguer. What does morphokinetics add to embryo selection and in-vitro fertilization outcomes? *Current Opinion in Obstetrics and Gynecology*, 27(3):193–200, 2015.

[5] Lorena Bori, Elena Paya, Lucia Alegre, Thamara Alexandra Viloria, Jose Alejandro Remohi, Valery Naranjo, and Marcos Meseguer. Novel and conventional embryo parameters as input data for artificial neural networks: an artificial intelligence model applied for prediction of the implantation potential. *Fertility and Sterility*, 114(6):1232–1241, 2020.

[6] Alejandro Chavez-Badiola, Adolfo Flores-Saiffe Farias, Gerardo Mendizabal-Ruiz, Rodolfo Garcia-Sanchez, Andrew J Drakeley, and Juan Paulo Garcia-Sandoval. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. *Scientific reports*, 10(1):1–6, 2020.

[7] Alejandro Chavez-Badiola, Adolfo Flores-Saiffe-Farías, Gerardo Mendizabal-Ruiz, Andrew J Drakeley, and Jacques Cohen. Embryo ranking intelligent classification algorithm (erica): artificial intelligence clinical assistant predicting embryo ploidy and implantation. *Reproductive BioMedicine Online*, 41(4):585–593, 2020.

[8] Fang Chen, Diane De Neubourg, Sophie Debrock, Karen Peeraer, Thomas D'Hooghe, and Carl Spiessens. Selecting the embryo with the highest implantation potential using a data mining based prediction model. *Reproductive Biology and Endocrinology*, 14(1):1–12, 2016.

[9] Fang Chen, Sophie Debrock, Karen Peeraer, Thomas D'hooghe, and Carl Spiessens. Selecting the embryo with the highest implantation potential using developmental and morphometric scoring. *Archives of Public Health*, 73:1, 2015.

[10] Giorgio Corani, Cristina Magli, Alessandro Giusti, Luca Gianaroli, and Luca Maria Gambardella. A bayesian network model for predicting pregnancy after in vitro fertilization. *Computers in biology and medicine*, 43(11):1783–1792, 2013.

[11] David K Gardner and Basak Balaban. Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'omics': is looking good still important? *MHR: Basic science of reproductive medicine*, 2016.

[12] David K Gardner, Michelle Lane, John Stevens, Terry Schlenker, and William B Schoolcraft. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertility and sterility*, 73(6):1155–1158, 2000.

[13] Eduardo Gazzo, Fernando Peña, Federico Valdéz, Arturo Chung, Claudio Bonomini, Mario Ascenzo, Marcelo Velit, and Ernesto Escudero. The kidscoretm d5 algorithm as an additional tool to morphological assessment and pgt-a in embryo selection: a time-lapse study. *JBRA assisted reproduction*, 24(1):55, 2020.

[14] Ruey-Shiang Guh, Tsung-Chieh Jackson Wu, and Shao-Ping Weng. Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes. *Expert Systems with Applications*, 38(4):4437–4449, 2011.

[15] Md Rafiul Hassan, Sadiq Al-Insaif, M Imtiaz Hossain, and Joarder Kamruzzaman. A machine learning approach for prediction of pregnancy outcome following ivf treatment. *Neural computing and applications*, 32(7):2283–2297, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[17] D Hlinka, B Kal'atova, I Uhrinova, S Dolinska, J Rutarova, J Rezacova, S Lazarovska, and M Dudas. Time-lapse cleavage rating predicts human embryo viability. *Physiological Research*, 61(5):513, 2012.

[18] Jan Holte, Lars Berglund, K Milton, C Garello, Gianluca Gennarelli, Alberto Revelli, and Torbjörn Bergh. Construction of an evidence-based integrated morphology cleavage embryo score for implantation potential of embryos scored and transferred on day 2 after oocyte retrieval. *Human Reproduction*, 22(2):548–557, 2007.

[19] Pegah Khosravi, Ehsan Kazemi, Qiansheng Zhan, Jonas E Malmsten, Marco Toschi, Pantelis Zisimopoulos, Alexandros Sigaras, Stuart Lavery, Lee AD Cooper, Cristina Hickman, et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ digital medicine*, 2(1):1–9, 2019.

[20] Mikkel F Kragh, Jens Rimestad, Jørgen Berntsen, and Henrik Karstoft. Automatic grading of human blastocysts from time-lapse imaging. *Computers in biology and medicine*, 2019.

[21] Mingzhao Li, Wanqiu Zhao, Wei Li, Xiaoli Zhao, and Juanzi Shi. Prognostic value of three pro-nuclei (3pn) incidence in elective single blastocyst-stage embryo transfer. *International journal of clinical and experimental medicine*, 8(11):21699, 2015.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[23] Yanhe Liu, Vincent Chapple, Katie Feenan, Peter Roberts, and Phillip Matson. Time-lapse deselection model for human day 3 in vitro fertilization embryos: the combination of qualitative and quantitative measures of embryo growth. *Fertility and sterility*, 105(3):656–662, 2016.

[24] Paweł Malinowski, Robert Milewski, Piotr Ziniewicz, Anna Justyna Milewska, Jan Czerniecki, and Sławomir Wołczyński. 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, 2014.

[25] Radwa Ali Mehanna. *Cell Culture*. BoD–Books on Demand, 2019.

[26] Marcos Meseguer, Javier Herrero, Alberto Tejera, Karen Marie Hilligsøe, Niels Birger Ramsing, and Jose Remohí. The use of morphokinetics as a predictor of embryo implantation. *Human reproduction*, 26(10):2658–2671, 2011.

[27] Robert Milewski, Jan Czerniecki, Agnieszka Kuczyńska, Bożena Stankiewicz, and Waldemar Kuczyński. Morphokinetic parameters as a source of information concerning embryo developmental and implantation potential. *Ginekologia polska*, 87(10):677–684, 2016.

[28] Seyed Abolghasem Mirroshandel, Fatemeh Ghasemian, and Sara Monji-Azad. Applying data mining techniques for increasing implantation rate by selecting best sperms for intra-cytoplasmic sperm injection treatment. *Computer methods and programs in biomedicine*, 137:215–229, 2016.

[29] American College of Obstetricians, Gynecologists, et al. Multiple gestation: Complicated twin, triplet, and high-order multifetal pregnancy, acog practice bullatin no. 56. *Obstet Gynecol*, 104:869–883, 2004.

[30] Evangelos G Papanikolaou, Efstratios M Kolibianakis, Herman Tournaye, Christos A Venetis, Human Fatemi, Basil Tarlatzis, and Paul Devroey. Live birth rates after transfer of equal number of blastocysts or cleavage-stage embryos in ivf. a systematic review and meta-analysis. *Human Reproduction*, 23(1):91–99, 2008.

[31] Bjørn Molt Petersen, Mikkel Boel, Markus Montag, and David K Gardner. Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on day 3. *Human reproduction*, 2016.

[32] Behnaz Raef and Reza Ferdousi. A review of machine learning approaches in assisted reproductive technologies. *Acta Informatica Medica*, 27(3):205, 2019.

[33] Arnaud Reignier, Jenna Lammers, Paul Barriere, and Thomas Freour. Can time-lapse parameters predict embryo ploidy? a systematic review. *Reproductive biomedicine online*, 2018.

[34] Kevin S Richter, Dee C Harris, Said T Daneshmand, and Bruce S Shapiro. Quantitative grading of a human blastocyst: optimal inner cell mass size and shape. *Fertility and sterility*, 76(6):1157–1167, 2001.

[35] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[36] Natalia Vladimirovna Saraeva, Natalia Vladimirovna Spiridonova, Marat Talgatovich Tugushev, Oksana Victorovna Shurygina, Sergey Igorevich Arabadzhyan, and Ivanova Olga Victorovna. Experience of using time lapse microscopy in the ivf program in patients with good ovarian reserve. *Gynecological Endocrinology*, 2019.

[37] Alpha Scientists. The istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Human reproduction*, 26(6):1270–1283, 2011.

[38] Lynette Scott, Ruben Alvero, Mark Leondires, and Bradley Miller. The morphology of human pronuclear embryos is positively related to blastocyst development and implantation. *Human reproduction*, 15(11):2394–2403, 2000.

[39] Shai Shalev-Shwartz, Yonatan Wexler, and Amnon Shashua. Shareboost: Efficient multiclass learning with feature sharing. *arXiv preprint arXiv:1109.0820*, 2011.

[40] Bruce S Shapiro, Said T Daneshmand, Forest C Garner, Martha Aguirre, and Shyni Thomas. 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, 2008.

[41] Jason E Swain. Could time-lapse embryo imaging reduce the need for biopsy and pgs? *Journal of assisted reproduction and genetics*, 30(8):1081–1090, 2013.

[42] Yun Tian, Wei Wang, Yabo Yin, Weizhou Wang, Fuqing Duan, and Shifeng Zhao. Predicting pregnancy rate following multiple embryo transfers using algorithms developed through static image analysis. *Reproductive biomedicine online*, 34(5):473–479, 2017.

[43] D Tran, S Cooke, PJ Illingworth, and DK Gardner. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Human Reproduction*, 34(6):1011–1018, 2019.

[44] Laura van Loendersloot, Sjoerd Repping, PMM Bossuyt, Fulco van der Veen, and Madelon van Wely. Prediction models in in vitro fertilization; where are we? a mini review. *Journal of advanced research*, 5(3):295–301, 2014.

[45] Lucinda L Veeck and Nikica Zaninovic. *An atlas of human blastocysts*. CRC Press, 2003.

Supplementary for Pseudo contrastive labeling for predicting IVF embryo developmental potential

I. Erlich[1,2], A. Ben-Meir[3,2], I. Har-vardi[4,2], J. Grifo[5], F. Wang[5], C. Mccaffrey[5], D. McCulloh[5], Y. Or[6], and L. Wolf[7]

[1]The Alexender Grass Center for Bioengineering, School of computer Science and Engineering, Hebrew University of Jerusalem, Israel
[2]Fairtilty Ltd., Tel Aviv, Israel
[3]IVF unit, Department of Obstetrics and Gynecology, Hadassah Medical Organization and Faculty of Medicine, Hebrew University of Jerusalem, Israel
[4]Fertility and IVF Unit, Department of Obstetrics and Gynecology, Soroka University Medical Center and the Faculty of Health Sciences Ben-Gurion University of the Negev, Beer-Sheva, Israel
[5]New York University Langone Prelude Fertility Center, New York, New York
[6]Fertility and IVF Unit, Obstetrics and Gynecology Division, Kaplan medical center
[7]The School of Computer Science, Tel Aviv University

A Additional Results

Fig. 1 and Fig. 2 are extended version of the corresponding paper figures, which contain all four datasets (for brevity, the paper only included the first two panesl of each).

Panel (c) in both figures addresses the task of identifying KIDp from discarded. The motivation for considering such a classifier, is to compare directly to a human embryologist. Embryologists perform the task of classification between discarded and non-discarded (i.e. KIDp+KIDn) embryos quite effectively.

The hardest discarded embryos are those that got to the blastulation stage, and yet were found to have poor morphology (such as CC graded embryos, using the Garnder method [1]). Those embryos can often be filtered out only after

the morphology is clear, at the late stages of blastulation. In panel (d) of both figures, we present identification of KIDp with the hardest discarded.

As can be seen in Fig. 1, both single-image pair models (C and E), outperform the proprietary day 3 KIDScore-D3 model, at all times for hard-discarded (d) and starting from 40 hours for discarded (c), despite the latter requiring 66 hours of continuous monitoring and manual annotations. Our initial model A is on par with KIDScore-D3 only late (106 hours), demonstrating the need to deploy the other steps suggested by our method.

Fig. 2 demonstrates that all pair models (C-F) outperform the initial models (A-B) which outperform the KIDScore-D3 model when predicting KID status across all age groups. The second integration pair model (F) outperforms the first pair integration model (D) across all ages. Further, even the latest single-image pair model (E) outperforms the first pair integration model (D) at most age groups, which empathises the efficiency of the second pair iteration in moving from a sequence-based decision to a single frame.

Interestingly, as can be seen from Fig. 2(a) KIDp vs. KIDn accuracy declines with age. KIDp vs. aneuploidy, discarded and hard-discarded (panels b-d) are more stable as a function of age. This is intuitive, since infertility of aged patients is often a result of the lower quality of older oocyte (age related infertility) [2], whereas younger infertility cases are often associated with other background causes. Thus, typically at aged patients, by selecting good embryo, the infertility cause is eliminated. Yet, for younger patients, selecting good embryo is often insufficient for a positive outcome. This, in turn, means that KIDn label is more noisy at younger ages. Aneuploid, discarded and hard-discarded tasks, are purely related to the embryo, thus not affected by this phenomena.

B Ablation Study

Table. 1 is an extended version of Table. 2 from the paper which includes additional models, which were trained using pseudo labels but with a the binary cross entropy loss instead of the pair loss.

Model G is analogous to model C in that it uses pseudo labels generated by model B, only a multiclass loss is used. Model H integrates model G. Model I is then trained using pseudo labels generated by H. The final ablation model is analog to model F in that it integrates the second pseudo labels model, in this case model I.

As can be seen, across all datasets, Model C is better than G, D is better than H, E is better than I, and finally, F is better than J. This indicates of the advantage of the pairwise approach.

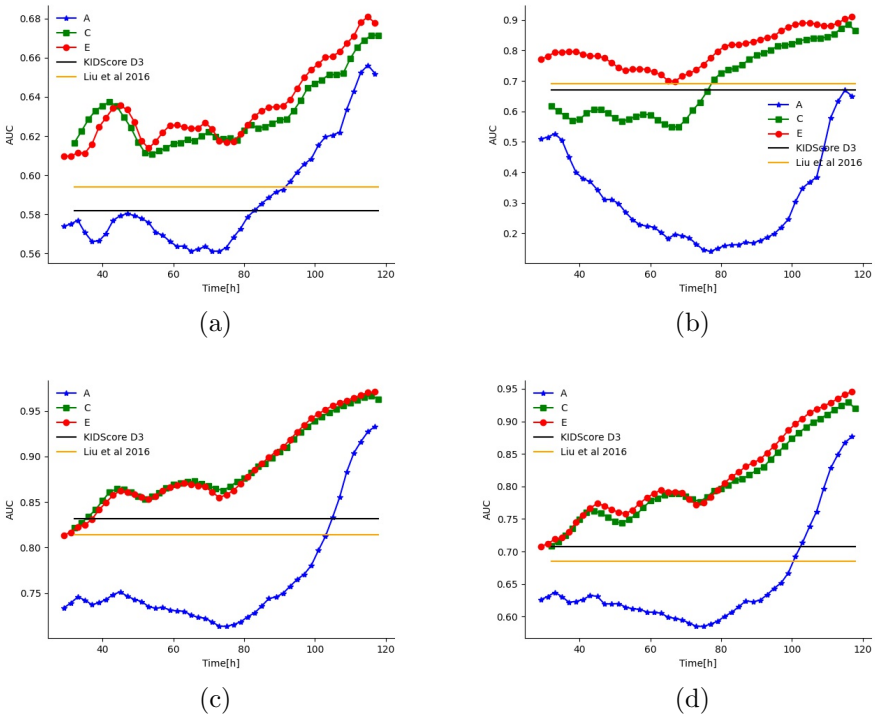


Figure 1: Prediction ability (AUC) as a function of time since fertilization (hours) for single-frame models. (a) KIDp vs. KIDn classification, (b) KIDp vs. aneuploid, (c) KIDp vs. discarded, (d) KIDp vs. hard discarded.

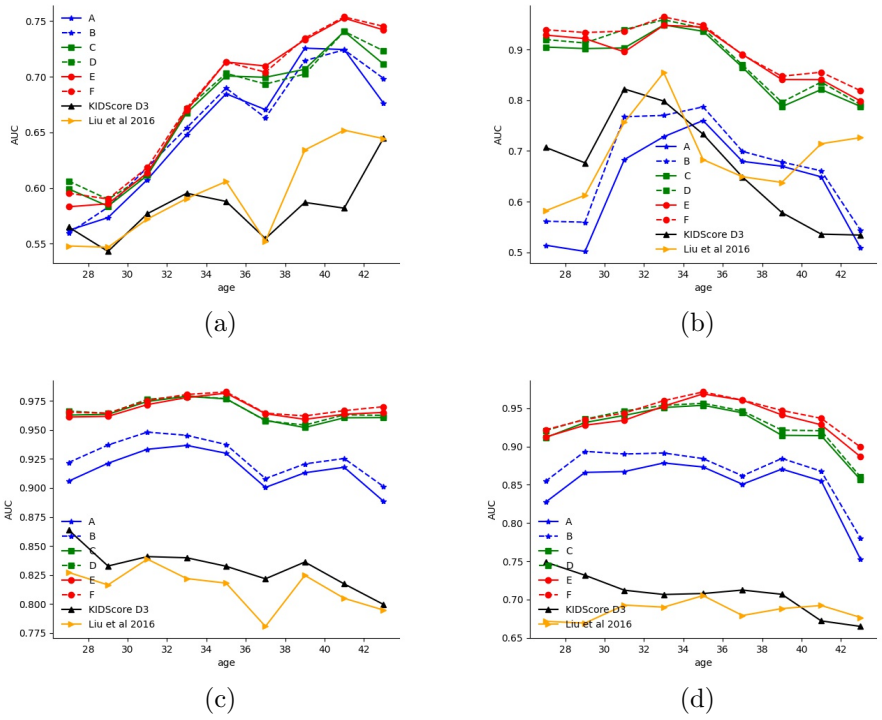


Figure 2: Prediction ability (AUC) as a function of oocyte age (years). (a) KIDp vs. KIDn classification, (b) KIDp vs. aneuploid, (c) KIDp vs. discarded, (d) KIDp vs. hard discarded.

Table 1: Classification results (AUC) for each model, including ablations models, for the different dataset (each denoted by the negative class) at the end of day 3 and at the end of day 5.

Model	Day	KIDn	Aneuplody	Discarded	Hard-Discarded
A (multi class model)	3	0.561	0.180	0.718	0.590
	5	0.653	0.633	0.916	0.849
B (integration of A)	5	0.656	0.67	0.927	0.868
C (pair model over B)	3	0.620	0.604	0.865	0.780
	5	0.669	0.873	0.965	0.924
D (integration of C)	5	0.671	0.885	0.967	0.929
E (pair model over D)	3	0.624	0.724	0.861	0.780
	5	0.678	0.890	0.967	0.935
F (integration of E)	5	**0.681**	**0.904**	**0.970**	**0.942**
G (multi class model, pseudo-label over B)	3	0.565	0.360	0.744	0.683
	5	0.665	0.801	0.944	0.904
H (integration of G)	5	0.664	0.819	0.949	0.913
I (multi class model, pseudo-label over H)	3	0.605	0.662	0.836	0.740
	5	0.669	0.878	0.961	0.924
J (integration of I)	5	0.670	0.898	0.965	0.933

References

[1] Gardner, D.K., Lane, M., Stevens, J., Schlenker, T., Schoolcraft, W.B.: Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. Fertility and sterility **73**(6), 1155–1158 (2000)

[2] Liu, K., Case, A., Cheung, A.P., Sierra, S., AlAsiri, S., Carranza-Mamane, B., Dwyer, C., Graham, J., Havelock, J., Hemmings, R., et al.: Advanced reproductive age and fertility. Journal of Obstetrics and Gynaecology Canada **33**(11), 1165–1175 (2011)

DISCUSSION AND CONCLUSION

In vitro fertilization (IVF) procedures have been used in clinical practice for the past 40 years. To this day, however, there are no reliable non-invasive methods available for identifying the small fraction of embryos that possess the highest potential to develop into a blastocyst that can be implanted and hopefully proceed to full-term delivery (Alpha Scientists,2011; Guerif et al., 2007). Common practice often involves the transfer of multiple embryos into the uterus, which in turn leads to multiple-embryo pregnancies and the risk of associated complications to both the newborn and the mother (Gleicher et al., 2016; Pinborg, 2005). The latter can be avoided by the transfer of a single embryo alone. However, traditional embryo selection methods, which are mostly based on morphological markers, are not predictive enough to allow routine single embryo transfer with reasonable implantation rates; therefore, new scoring tools are needed.

The emergence of TLI has led to a range of classifications and grading systems to assess embryo quality. Several markers related to the pronuclei stage (Lynette et al. ,2000, Scott et al., 2000, Aguilar et al., 2014, Li et al., 2015); the cleavage stage (Holte et al. 2007, Meseguer et al. 2011, Yang et al., 2015, Gardner et al., 2016) and the blastocyst stage (Gardner et al., 2000, Hassan et al. 2018, Bori et at., 2020,Chavez-Badiola et al., 2020) have been suggested to predict embryo viability. Saraeva et al. (2019) demonstrate a correlation between spontaneous blastocyst collapse and poor pregnancy outcomes. However, these have not always been translated into quantitative implantation grading. Further, they were based on a biomarker appearing at a specific time point and ignoring the rest of the TLI stream. My study (chapter 1) was inspired by this biomarkers approach, proposing a new training scheme that identifies golden markers over short time periods called packets. We propose an adaptive weighted hinge loss as well as a time-dependent architecture with shared weights between times for training Convolutional Neural Network (CNN). The CNN is trained to find biomarkers to predict embryo viability at only some of the packets, while minimizing the amount of "fake markers". Using the same approach as Lin et al., (2017), the weights are tuned across packets of the same embryo according to whether one of the packets is a golden marker. Seeing as another packet already located a golden marker, the

weight of the rest of the packets becomes lighter. A multi-frame fusion of the packets' markers creates a continuous score for the embryo, according to all data available at that point in time, which overcomes the shortcomings of the local markers approach described previously.

Our results show a monotonically increasing Area under the Curve (AUC) up to 0.85 at 90 hours for blastulation formation prediction ,compared to Milewski et al. (2016) consisting of an AUC of up 0.76 over the same time period and data, and AUCs ranging up to 0.76 for implantation prediction of dataset including Day 3 and Day 5 transfers, compared to AUC of 0.58 achieved by the commercial KIDSCORE-D3.

Since most embryos are not transferred back to the uterus, the lack of labeled data hampers the development of embryo grading systems. In addition, implantation failures are often caused by non-embryonic reasons, which result in a false negative label. For this, a different learning scheme or end point label is needed. In our work (chapter 4), we address both problems simultaneously by presenting a pseudo-contrastive label scheme that compares paired embryos instead of singletons. The task of predicting implantation outcome is therefore replaced with the task of rightly rank all the pairs inside a multi-batch of paired embryos. During the training process, a cascade of models is iteratively trained, where the grand-truth label is determined by the score of the previous model. The first model in the cascade is the vanilla fully supervised model. Our new presented framework enables us to increase training amounts by six fold and more (Tran et al, 2019). We demonstrated the effectiveness of the pseudo-contrastive labels scheme when compared to both supervised method and previously reported methods, including the commercial KIDScore-D3 system, and a group of eight senior professionals over all times and patient ages. Specifically, the pseudo-contrastive labels model achieved an AUC of 0.699, compared to the average AUC of 0.631 achieved by the senior embryologists' group, and 0.594 and 0.582 achieved by Liu et al. al. (2016) and the commercial KIDScore-D3, respectively, over a testset of 1017 Day 5 transfers, consisting of 380/637 implanted/ non implanted embryos.

Both of the above mentioned studies do not explicitly consider the overall morphokinetic development profile of the embryo, which is considered to be one of the major advantages that can be gained using TLI. Similarly, since the emergence of TLI, most studies for grading embryos can be roughly divided into two main categories: morphological based, and morphokinetic based. Most morphokinetic models consist of a set of hand peaked rules involving specific events timings, typically till eight cells/cleavage stage (Hlinka et al., 2012, Basile et al., 2014, Meseguer et al.,

2014, Liu et al., 2016, Milewski et al., 2016, Pirkevi et al, 2016). By contrast, although few morphology models were suggested for the cleavage stage (Alpha Scientists, 2011, Nasiri et al., 2015), the majority of studies aimed for Day 5, at which morphology was found to be the most effective, as the embryo enters the late blastocyst stages (Garnder et al. 2000, Papanikolaou et a., 2006, Khosravi et al. 2019, Kragh et al. 2019, Bori et al. 2020, Chavez-Badiola et al., 2020, VerMilyea et al., 2020). This is also reflected in clinical practice, where speed and overall development path are mainly used during the early developmental stages, and morphological scales are applied during the later development stages. However, there was no previous study that compares both methodologies over the same test set, and comprehensively assesses the information available in these channels. The latter doesn't remain theoretical, but rather concerns practical questions such as which embryo should be preferred: a well developed embryo ending up with only fair morphology or a top graded embryo by its final morphology, yet with a less favorable developmental path. Here, we present two CNN models, one for each information category, that continuously evaluate embryo quality over time. For morphokinetics, we present a scheme consisting of two stages. First, we estimate 15 morphokinetic starts by employing a CNN over each individual frame in the TLI data, followed by operating dynamic programming to find the most likely morphokinetic path. The second is a CNN trained to predict implantation probability using the estimated morphokinetic timings as inputs. For morphology, we used our previously described model, based on the pseudo-contrastive labels framework. The amount of information exhibited by each model as a function of time is then compared between the two models across time. We observed monotonically increasing AUCs for both classifiers over time. Our findings indicate different trends in terms of information. Morphokinetic information appears already at an early stage, while morphological information is acquired much later, around 90 hours post fertilization, which corresponds to the start of the blastocyst stage. From that point on, the morphology model bypasses the morphokinetic model, which is well aligned with both the practical way of working, as well as the volume of studies where morphokinetic mostly applies at early stages, and morphology is primarily aimed for the late blastocyst stages (Kragh et al. 2019, Bori et al. 2020, Chavez-Badiola et al., 2020, VerMilyea et al., 2020). As a final step, we propose a joint CNN model with morphokinetic and morphology pretrained models, followed by concatenate and prediction dense layers. The joint morphokinetic - morphology model is then compared across time to both individual models, and is shown to outperform both, at all times.

In the last phase of our work, we observed a discrepancy between the way that embryologists choose which embryos to transfer in daily care, and the majority of the reported studies automating embryo selection with machine learning. Specifically, embryologists split the task into two subtasks. The first step is ranking the sibling embryos within the cohort to identify the best candidates for transfer. Then, after considering additional factors pertinent to the specific patient, such as age, weight, number of previous failed attempts to conceive, endometrium conditions, and other factors, a decision is made regarding the number of embryos to jointly transfer based on the ranking order from top to bottom. However, most studies that rely on machine learning replace these two steps by the single step of predicting implantation outcomes of the subgroup of transferred embryos and then applying the same scoring scheme for the purpose of ranking. The underlying assumption is that a model that could differentiate between implanted and non-implanted embryos, is certain to classify implanted embryos from non-transferred embryos, as those who were not transferred are likely to possess a worse appearance than those who were transferred. Here, we challenge this axiom. We differentiate between the task of ranking embryos in a cohort and predicting implantation probability for each individual embryo, taking into account the overall clinical context. Similarly, we differentiate between factors that are unique to each individual embryo - embryo-specific - and factors shared by all the individuals within a group of siblings - cohort-specific. The former includes factors such as overall competence, pronuclei characteristics (Aguilar, et al., 2014; Scott, Alvero, Leondires, & Miller, 2000; Li, Zhao, Li, Zhao, & Shi, 2015), morphokinetic temporal events (Petersen, Boel, Montag, & Gardner, 2016; Liu, Chapple, Feenan, Roberts, & Matson, 2016), genetic ploidy status (Guh et al., 2011, Swain et al., 2013) and morphological markers (Garnder et al. 2000, Kragh et al. 2019, Bori et al. 2020, Chavez-Badiola et al., 2020). The latter includes clinical factors such as the age of the patient and the oocyte, the patient's weight, non-receptive endometrium, as well as other background factors. Despite a consensus that these clinical properties are highly correlated with implantation outcomes, it is by definition impossible for them to contribute to the task of ranking sibling embryos, given that they are shared by all embryos within a cohort. For this, we redefine the optimization task as ranking embryos by their viability, and consider only embryo-specific factors. Specifically, we demonstrate the difference between both optimization tasks by comparing a model learned to predict implantation with the oocyte age, and a model trained without, over different negative groups. Our findings indicate that despite the positive impact of oocyte age on predicting

implantation, the ability to distinguish between implanted embryos and other negatives, generally considered as easier than predicting the transferred group, deteriorates compared to the model trained without the oocyte age. We Further analyze the implications of learning from implantation outcomes, where a negative outcome does not necessarily signify a bad embryo. We found that this label noise adversely affects the ability to differentiate between viable and non-viable embryos, even for implantation prediction, in comparison to training with trivial, but clean, embryos, as in discarded embryos. Furthermore, we discuss the significance of training with trivial negatives such as underdeveloped discarded and discarded that developed to late blastocyst stages, but were discarded due to poor morphology. The former was as good as learning with non-implanted embryos (KID-N), and the latter outperformed KID-N on all test sets. Further, our results indicated that learning from both KID-N and discarded examples generated the best results on all negative test sets. Thus, besides reflecting the clinicians' practice for selecting which embryos to transfer in an IVF cycle, another benefit of this research is that it provides evidence for learning with all negative data, including those more trivial discarded, to improve classification results across both tasks.

Through bringing attention to all of the aspects discussed in this dissertation, we hope to provide a basis for future studies and to improve the quality of IVF sessions.

REFERENCES

ACOG Practice Bulletin #56 (2004). "Multiple gestation: complicated twin, triplet, and high- order multifetal pregnancy." *Obstet Gynecol* 104:869-883.

Aguilar J, Motato Y, Escribá MJ, Ojeda M, Muñoz E, Meseguer M. The human first cell cycle: impact on implantation. Reprod Biomed Online. 2014 Apr;28(4):475-84. doi: 10.1016/j.rbmo.2013.11.014. Epub 2013 Dec 11. PMID: 24581982.

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 Assist Reprod* 22(3):205–214.

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

Aparicio-Ruiz, B., Romany, L., Meseguer, M. (2018). "Selection of preimplantation embryos using time-lapse microscopy in in vitro fertilization: State of the technology and future directions." *Birth Defects Research* 110:648–653.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

Avendi, M. R., Kheradvar, A., Jafarkhani, H. (2016). "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI." *Medical Image Analysis* 30:108–119.

Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H. (2015). "Chest pathology detection using deep learning with non-medical training." *Biomedical Imaging (ISBI), IEEE 12th International Symposium.* pp. 294–297.

Basile, N, Caiazzo, M, Meseguer, M. (2015). "What does morphokinetics add to embryo selection and in-vitro fertilization outcomes?" *Current Opinion in Obstetrics and Gynecology* 3:193–200.

Basile, N. Meseguer, M. (2014). "Time-lapse technology: Evaluation of embryo quality and new markers for embryo selection." *Expert Review of Obstetrics and Gynecology,*7.

Blake, D.A., Farquhar, C.M., Johnson, N., Proctor, M. (2007). "Cleavage stage versus blastocyst stage embryo transfer in assisted conception." *Cochrane Database Syst. Rev*.,

Bori L, Paya E, Alegre L, Viloria TA, Remohi JA, Naranjo V, Meseguer M. Novel and conventional embryo parameters as input data for artificial neural networks: an artificial intelligence model applied

for prediction of the implantation potential. Fertil Steril. 2020 Sep 8:S0015-0282(20)30777-9. doi: 10.1016/j.fertnstert.2020.08.023. Epub ahead of print. PMID: 32917380.

Brosch, T., Tam, R., and Alzheimer's Disease Neuroimaging Initiative. (2013). "Manifold learning of brain MRIs by deep learning." *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 633-640). (Berlin, Heidelberg: Springer).

Chavez-Badiola, A., Flores-Saiffe Farias, A., Mendizabal-Ruiz, G. *et al.* Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. *Sci Rep* **10,** 4394 (2020). https://doi.org/10.1038/s41598-020-61357-9

Chavez-Badiola, Alejandro & Flores Saiffe, Adolfo & Mendizabal-Ruiz, Gerardo & Drakeley, Andrew & Cohen, Jacques. (2020). Embryo Ranking Intelligent Classification Algorithm (ERICA), an artificial intelligence clinical assistant with embryo ploidy and implantation predicting capabilities. Reproductive BioMedicine Online. 10.1016/j.rbmo.2020.07.003.

Chen, F., De Neubourg, D., Debrock, S., Peeraer, K., D'Hooghe, T., Spiessens, C. (2016). "Selecting the embryo with the highest implantation potential using a data mining based prediction model." *Reprod Biol Endocrinol*. 14:10.

Chen F, De Neubourg D, Debrock S, Peeraer K, D'Hooghe T, Spiessens C. Selecting the embryo with the highest implantation potential using a data mining based prediction model. Reproductive Biology and Endocrinology. 2016; 14: 10. doi:10.1186/s12958-016-0145-1.

Clift, D., Schuh, M. (2013). "Restarting life: fertilization and the transition from meiosis to mitosis." *Nature Reviews. Molecular Cell Biology* 14(9):549–562.

Collins, J., van Knippenberg, B., Ding, K., Kofman, A. (2019). "Time-Lapse Microscopy";in Radwa Ali Mehanna, Cell Culture, Intechopen, 2019.

Corani G, Magli C, Giusti A, Gianaroli L, Gambardella LM. A Bayesian network model for predicting pregnancy after in vitro fertilization. Computers in biology and medicine. 2013; 43: 1783-1792. doi:10.1016/j.compbiomed.2013.07.035.

Eskew, A.M,, Jungheim, E.S. (2017). "A history of developments to improve in vitro fertilization." *Mo Med* 114(3):156–159.

Frizzi, S., Kaabi, R., Bouchouicha, M., Ginoux, J.-M., Moreau, E., Fnaiech, F. (2016). "Convolutional neural network for video fire and smoke detection." *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*, pp. 877–882.

Gardner, D., Kelley, R. (2017). "Impact of the IVF laboratory environment on human preimplantation embryo phenotype." Journal of Developmental Origins of Health and Disease, 8(4), 418–435.

Gardner, D.K., Balaban, B. (2016). "Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': is looking goodstill important? *Molecular Human Reproduction* 22(10):704–718.

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. *Fertil Steril.* 73(6):1155–1158.

Gardner, D.K., Meseguer, M., Rubio, C., Treff, N.R. (2015). "Diagnosis of human preimplantation embryo viability." *Hum. Reprod Update* 21:727–747.

Gerris, J.M.R. (2005). "Single embryo transfer and IVF/ICSI outcome: a balanced appraisal." *Human Reproduction Update* 11(2):105,121.

Gers, F. A.; Schmidhuber, E. (November 2001). "LSTM recurrent networks learn simple context-free and context-sensitive languages" (PDF). IEEE Transactions on Neural Networks. 12 (6): 1333–1340. doi:10.1109/72.963769. ISSN 1045-9227. PMID 18249962.

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):25.

Glujovsky, D., Farquhar, C., Quinteiro, Retamar, A.M., Alvarez Sedo, CR, Blake, D. (2016). Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology. *Cochrane Database of Systematic Reviews*, 6.

Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. (2009). "A Novel Connectionist System for Improved Unconstrained Handwriting Recognition" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 31 (5): 855–868. CiteSeerX 10.1.1.139.4502. doi:10.1109/tpami.2008.137. PMID 19299860. S2CID 14635907.

Guerif, F. et al. (2007). Limited value of morphological assessment at days 1 and 2 to predict blastocyst, development potential: a prospective study based on 4042 embryos. *Hum Reprod* 22, 1973-1981.

Guh RS, Wu TCJ, Weng SP. Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes. Expert Systems with Applications. 2011; 38: 4437-4449. doi:https://doi.org/10.1016/j.eswa.2010.09.112.

Hassan MR, Al-Insaif S, Hossain MI, Kamruzzaman J. A machine learning approach for prediction of pregnancy outcome following IVF treatment. Neural Computing and Applications. 2018: 1-15. doi:10.1007/s00521-018-3693-9

Hatırnaz, Ş, Kanat Pektaş M. (2017). Day 3 embryo transfer versus day 5 blastocyst transfers: A prospective randomized controlled trial. *Turk J Obstet Gynecol* 14(2):82–88.

He, K., Gkioxari, G., Dollar, P., Girshick, R. (2017). "Mask R-CNN." *Proceedings of the IEEE International Conference on Computer Vision*.

Hlinka, D. et al. (2012). Time-lapse cleavage rating predicts human embryo viability.*Physiol Res* 61:513–525.

Holte J, Berglund L, Milton K, Garello C, Gennarelli G, Revelli A, et al. Construction of an evidence-based integrated morphology cleavage embryo score for implantation potential of embryos scored and transferred on day 2 after oocyte retrieval. Human reproduction (Oxford, England). 2007; 22: 548-557. doi:10.1093/humrep/del403.

Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences of the United States of America. 1982 Apr;79(8):2554-2558. DOI: 10.1073/pnas.79.8.2554.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andretto, M., and Adam, H.(2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *ArXiv. 1704.04861*.

Jarvis, G.E. (2016). "Early embryo mortality in natural human reproduction: What the data say." *F1000Res*. 5:2765.

Jordan, Michael I. (May 1986). Serial order: a parallel distributed processing approach. Tech. rep. ICS 8604. San Diego, California: Institute for Cognitive Science, Universityof California

K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

Kattal, N., Cohen J. and Barmat, L.J. (2008). "Role of coculture in human in vitro fertilization: a meta-analysis." *Fertility and Sterility* 90(4): 1069–1076**.**

Khalaf, Y. (2013). "Cassandra's prophecy: why we need to tell the women of the future about age-related fertility decline and 'delayed' childbearing." *Reprod BioMed Online* 27,10.

Khosravi, P., Kazemi, E., Zhan, Q. *et al.* Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *npj Digit. Med.* **2,** 21 (2019). https://doi.org/10.1038/s41746-019-0096-y

Kobayashi, T., Akinori, H., Kurita, T. (2008). "Selection of histograms of oriented gradients features for pedestrian detection." *Neural Information Processing*, 4985:598–607.

Kragh MF, Rimestad J, Berntsen J, Karstoft H. Automatic grading of human blastocysts from time-lapse imaging. Comput Biol Med. 2019 Dec;115:103494. doi: 10.1016/j.compbiomed.2019.103494. Epub 2019 Oct 15. PMID: 31630027.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks." *Advances in Neural Information Processing Systems,* pp. 1097–1105.

Larsson, G., Maire, M., Shakhnarovich, G. (2016). "FractalNet: ultra-deep neural networks without residuals," *arXiv 1605. 07648,* 1–11.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE,* 86(11): 2278-2324.

Li M, Zhao W, Li W, Zhao X, Shi J. Prognostic value of three pro-nuclei (3PN) incidence in Outcome of embryo mixed transfer cycles 11005 Int J Clin Exp Med 2017;10(7):11001-11005 elective single blastocyst-stage embryo transfer. Int J Clin Exp Med 2015; 8: 21699-20702

Li, Xiangang; Wu, Xihong (2014-10-15). "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition". arXiv:1410.4281 [cs.CL].

Lin Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection." In *Proceedings of the IEEE international conference on computer vision*, pp. 2980-2988. 2017.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L. (2017). "SphereFace: Deep Hypersphere Embedding for Face Recognition." *arXiv preprint arXiv:1704.08063*.

Lynette Scott, Ruben Alvero, Mark Leondires, Bradley Miller, The morphology of human pronuclear embryos is positively related to blastocyst development and implantation, *Human Reproduction*, Volume 15, Issue 11, November 2000, Pages 2394–2403, https://doi.org/10.1093/humrep/15.11.2394

M. Sundermeyer, R. Schlüter, and H. Ney. (2012). "LSTM Neural Networks for Language Modeling", in *Interspeech-2012*, 194-197.

Malinowski P, Milewski R, Ziniewicz P, Milewska AJ, Czerniecki J, Wołczyński S. The use of data mining methods to Predict the Result of Infertility Treatment Using the IVF ET Method. Studies in Logic, Grammar and Rhetoric. 2014; 39: 67-74. doi:10.2478/slgr-2014-0044.

Malmsten, Jonas / Zaninovic, Nikica / Zhan, Qiansheng / Rosenwaks, Zev / Shan, Juan Automated cell division classification in early mouse and human embryos using convolutional neural networks 2021 *Neural Computing and Applications* , Vol. 33, No. 7 Springer p. 2217-2228

Mauri, A.L., Petersen, C.G., Baruffi, R.L., and Franco, J.G. Jr. (2001). "A prospective, randomized comparison of two commercial media for ICSI and embryo culture." *J Assist Reprod Genet*. 18:378–381.

McLernon, D. J., Maheshwari, A., Lee, A., Bhattacharya, S. (2016). "Cumulative live birth rates after one or more complete cycles of IVF: a population-based study of linked cycle data from 178,898 women." *Human Reproduction* 31(3):572–581.

Medical Advisory Secretariat (2006). "In vitro fertilization and multiple pregnancies: an evidence-based analysis." *Ontario health technology assessment series* 6(18):1–63.

Meseguer, M. et al. (2011). "The use of morphokinetics as a predictor of embryo implantation." *Hum Reprod.* 26:2658-2671.

Meseguer, Marcos. (2020). Novel and conventional embryo parameters as input data for artificial neural networks: an artificial intelligence model applied for prediction of the implantation potential. Fertility and sterility. 10.1016/j.fertnstert.2020.08.023.

Mikolov, T., Karafiát, M., Burget, L., Černocky, J., Khudanpur, S. (2010). “Recurrent neural network based language model,” in *Eleventh Annual Conference of the International Speech Communication Association*.

Milewski, R., Czerniecki, J., Kuczyńska, A., Stankiewicz, B., Kuczyński, W. (2016b). "Morphokinetic parameters as a source of information concerning embryo developmental and implantation potential." *Ginekologia Polska* 87:677–684.

Mirroshandel SA, Ghasemian F, Monji-Azad S. Applying data mining techniques for increasing implantation rate by selecting best sperms for intra-cytoplasmic sperm injection treatment. Computer methods and programs in biomedicine. 2016; 137: 215-229. doi:10.1016/j.cmpb.2016.09.013.

Nakahara, T. et al. "Evaluation of the safety of time-lapse observations for human embryos." *J Assist Reprod Genet* 27:93–96.

Nasiri, N., & Eftekhari-Yazdi, P. (2015). An overview of the available methods for morphological scoring of pre-implantation embryos in in vitro fertilization. *Cell Journal (Yakhteh)*, *16*(4), 392.

Natalia Vladimirovna Saraeva, Natalia Vladimirovna Spiridonova, Marat Talgatovich Tugushev, Oksana Victorovna Shurygina, Sergey Igorevich Arabadzhyan & Ivanova Olga Victorovna (2019) Experience of using time lapse microscopy in the IVF program in patients with good ovarian reserve, Gynecological Endocrinology, 35:sup1, 15-17, DOI: 10.1080/09513590.2019.1632090

Ng, P.C., Henikoff, S. (2003). "SIFT: Predicting amino acid changes that affect protein function." *Nucleic Acids Res*. 31(13):3812–3814.

Niakan, K. K., Han, J., Pedersen, R. A., Simon, C., Pera, R. A. (2012). "Human pre-implantation embryo development." *Development* (Cambridge, UK), 139(5), 829–841.

Niederberger, C., Pellicer, A., Cohen, J., Gardner, D.K., Palermo, G.D., O'Neill, C.L., et al. (2018). "Forty years of IVF." *Fertil Steril*. 110:185–324.

O. Tadmor, T. Rosenwein, S. Shalev-Shwartz, Y. Wexler, and A. Shashua. Learning a metric embedding for face recognition using the multibatch method. In NIPS, 2016.

Ombabi, Abubakr & Ouarda, Wael & Alimi, Adel. (2020). Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. Social Network Analysis and Mining. 10. 10.1007/s13278-020-00668-1.

Ombelet, W., Cooke, I., Dyer, S., Serour, G., Devroey, P. (2008a). "Infertility and the provision of infertility medical services in developing countries." *Human Reprod Update* 14:605–621.

Papanikolaou, E.G., Camus, M., Kolibianakis, E.M., Van Landuyt, L., Van Steirteghem, A.,Devroey, P. (2006). "In vitro fertilization with single blastocyst-stage versus single cleavage-stage embryos." *N. Engl. J. Med*. 354:1139–1146.

Papanikolaou, E.G., Kolibianakis, E.M., Tournaye, H., Venetis, C.A., Fatemi, H., Tarlatzis,B., Devroey, P. (2008). "Live birth rates after transfer of equal number of blastocysts or cleavage-stage embryos in IVF. A systematic review and meta-analysis." *Hum Reprod*. 23: 91–99.

Park H, Lee HJ, Kim HG, Ro YM, Shin D, Lee SR, Kim SH, Kong M. Endometrium segmentation on transvaginal ultrasound image using key-point discriminator. Med Phys.2019 Sep;46(9):3974-3984. doi: 10.1002/mp.13677. Epub 2019 Jul 31. PMID: 31230366.

Pinborg, A. (2005). "IVF/ICSI twin pregnancies: risks and prevention." *Human Reproduction Update* 11(6):575–59.

Pirkevi Çetinkaya, C., Kahraman, S. (2016). "Morphokinetics of embryos – where are we now?" *Journal of Reproductive Biotechnology and Fertility*, 5.

Pirkevi Çetinkaya, C., Kahraman, S. (2016). "Morphokinetics of embryos − where are we now?" *Journal of Reproductive Biotechnology and Fertility*, August

Raef, Behnaz & Ferdousi, Reza. (2019). A Review of Machine Learning Approaches in Assisted Reproductive Technologies. Acta Informatica Medica. 27. 205 10.5455/aim.2019.27.205-211.

Reignier, A. et al. (2018). "Can time-lapse parameters predict embryo ploidy? A systematic review."Reproductive BioMedicine Online 36(4):380–387.

Richter, K.S., Harris, D.C., Daneshmand, S.T., Shapiro, B.S. (2001). "Quantitative grading of a human blastocyst: optimal inner cell mass size and shape." *Fertil Steril*. 76:1157–1167.

Ronneberger, O., Fischer, P., Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241).

Rubio I, Kuhlmann R, Agerholm I, Kirk J, Herrero J, Escribá MJ, Bellver J, Meseguer MFertil Steril. 2012 Dec; 98(6):1458-63. Limited implantation success of direct-cleaved human zygotes: a time-lapse study.

Rumelhart, David E; Hinton, Geoffrey E, and Williams, Ronald J (Sept. 1985). Learning internal representations by error propagation. Tech. rep. ICS 8504. San Diego, California: Institute for Cognitive Science, University of California.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S. Ziheng, H., Karpathy, A., Khosla, A., Bernstein, M. Berg, A., Fei-Fei, Li.. (2015). "Imagenet large scale visual recognition challenge." *International Journal of Computer Vision*, 115(3):211–252.

Sak, Hasim; Senior, Andrew; Beaufays, Francoise (2014). "Long Short-Term Memory recurrent neural network architectures for large scale acoustic modeling" (PDF). Archived from the original (PDF) on 2018-04-24.

Schroff, F., Kalenichenko, D., Philbin, J. (2015). "Facenet: A unified embedding for face recognition and clustering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 815-823.

Segal S, Casper RF. Progesterone supplementation increases luteal phase endometrial thickness and oestradiol levels in in-vitro fertilization. Hum Reprod. 1992 Oct;7(9):1210-3. doi: 10.1093/oxfordjournals.humrep.a137828. PMID: 1478999.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735–1780. DOI:https://doi.org/10.1162/neco.1997.9.8.1735

Shalev-Shwartz, S., Shammah, S., Shashua, A. (2016). "Safe, multi-agent, reinforcement learning for autonomous driving." *arXiv preprint arXiv:1610.*03295.

Shalev-Shwartz, S., Wexler, Y., and Shashua, A. (2011). "Shareboost: Efficient multiclass learning with feature sharing." *Proc. NIPS.*

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." *Fertil Steril.* 90:302–309.

Suk, H. I., Lee, S. W., Shen, D., Alzheimer's Disease Neuroimaging Initiative. (2014). "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis." *NeuroImage* 101:569–582.

Sundvall, Linda / Ingerslev, Hans Jakob / Breth Knudsen, Ulla / Kirkegaard, Kirstine Inter-and intra-observer variability of time-lapse annotations 2013 *Human Reproduction* , Vol. 28, No. 12 Oxford University Press p. 3215-3221

Swain, J.E. (2013). "Could time-lapse embryo imaging reduce the need for biopsy and PGS?" *J Assist Reprod Genet*. 30(8):1081–1090.

Swain, J.E. et al. (2016). "Optimizing the culture environment and embryo manipulation to help maintain embryo developmental potential." *Fertility and Sterility* 105(3):571 – 587.

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*, 34(06):1011–1018.

Van den Abbeel, E., Balaban, B., Ziebe, S., Lundin, K., Cuesta, M. J., Klein, B. M., Helmgaard, L., Arce, J. C. (2013). "Association between blastocyst morphology and outcome of single-blastocyst transfer." *Reprod Biomed Online* 27:353–361.

van Loendersloot L, Repping S, Bossuyt PM, van der Veen F, van Wely M. Prediction models in in vitro fertilization; where are we? A mini review. J Adv Res. 2014 May;5(3):295-301. doi: 10.1016/j.jare.2013.05.002. Epub 2013 May 9. PMID: 25685496; PMCID: PMC4294714.

Veeck, L. L. & Zaninovic, N. An Atlas of Human Blastocysts, Vol. 286 (CRC Press, Taylor & Francis, 2003).

VerMilyea, M., Hall, J. M. M., Diakiw, S. M., Johnston, A., Nguyen, T., Perugini, D., ... & Perugini, M. (2020). Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. *Human Reproduction*, *35*(4), 770-784.

Wahab, N., Khan, A., Lee, Y. S. (2019). "Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images." *Microscopy* 68(3):216–233.

Wang, X., Du, M., Guan, Y., Wang, B., Zhang, J., Liu, Z. (2017). "Comparative neonatal outcomes in singleton births from blastocyst transfers or cleavage-stage embryo transfers: a systematic review and meta-analysis. *Reproductive biology and endocrinology* 15(1):36.

Wu Y, Gao X, Lu X, et al. Endometrial thickness affects the outcome of in vitro fertilization and embryo transfer in normal responders after GnRH antagonist administration. *Reprod Biol Endocrinol*. 2014;12:96. Published 2014 Oct 9. doi:10.1186/1477-7827-12-96

Yang, S. T. et al. (2015). "Cleavage pattern predicts developmental potential of day 3 huma embryos produced by IVF." *Reproductive BioMedicine Online* 30(6):625–634.

Zhu, X., Goldberg, A. B. (2009). Introduction to semi-supervised learning (San Rafael, CA: Morgan & Claypool).

תקציר

הפריה חוץ גופית (IVF) כוללת סדרה של הליכים בהם ביציות בוגרות נשאבות מהשחלות ומופרות על ידי זרע במעבדה. הביציות המופרות (עוברים) מתורבתות לתקופה של מספר ימים (בדרך כלל 3-6) ולאחר מכן מועברות לרחם. מחזור מלא של IVF בדרך כלל אורך כשלושה שבועות או יותרIVF היא הצורה היעילה ביותר של טיפולי פוריות. עם זאת, טיפולי הפריה חוץ גופית אינם יעילים, יקרים, ממושכים ומטילים נטל רגשי על ההורים לעתיד.

בפרט, בחירת העוברים בעלי פוטנציאל להשרשה הטוב ביותר נותרה משימה מאתגרת. למרות כמה עשרות שנים של פרקטיקה קלינית, אין שיטה לא פולשנית אמינה לזהות את החלק הקטן של עוברים שיש להם את הפוטנציאל הגבוה ביותר להתפתח לבלסטוציסט, שאחר כך ניתן להחזיר אותו לרחם בתקווה להמשך הריון. לפיכך, כדי לשמור על אחוזי הריון סבירים, הנוהג הנוכחי כרוך בהעברת עוברים מרובים לרחם. התוצאה היא סיבוכים קליניים וסיכונים בריאותיים ליילוד ולאם, עם הריונות מרובי עוברים (מעל 10% מההריונות חוץ גופיים הם תאומים או יותר).

כדי לשפר את שיעורי ההריון של טיפול IVF, ולהימנע מסיבוכים הן לאם והן לעובר, יש לזהות ולהחזיר רק את העוברים בעלי הסיכוי הגבוה ביותר להשרשה. לשם כך הוצעו מספר שיטות להערכת עוברים. רוב השיטות עד כה התייחסו לתמונות דו -ממדיות בלבד של העובר שצולמו על ידי מצלמה על גבי מיקרוסקופ רגעים לפני ההעברה.

הכנסת מערכות ניתור עוברי בזמן בשנים האחרונות מספקת הדמיה מתמשכת של העוברים תוך שמירה עליהם בתנאי תרבות אופטימליים. עם זאת, רק מחקרים בודדים נערכו על היישומים הפוטנציאליים של הדמיה תלת -ממדית (בזמן) להערכת עוברים ולזיהוי המשתלמים ביותר. אלה מבוססים על תיוג ידני, הכולל עומס עבודה משמעותי ומומחיות אנושית, מוגבלים לשלבי התפתחות מאוחרים יותר ודיווחו רק על הצלחה מוגבלת.

עבודת גמר זו מציעה מערכת דרוג עוברים אוטומטית אשר תספק סטנדרטיזציה לדירוג מודלים. ההיבטים השונים של מטרה זו נבחנים היטב. בשלב הראשון אנו מבחינים בחיזוי השרשה ודירוג העוברים, ששניהם בדרך כלל נתפסים כמשימה אחת. בנוסף להשוואת מודלים שונים על ידי AUC, אנו מציעים להשתמש במדד חדש בשם זמן להשגת הריון. בעזרת מדדים אלה, אנו מדגימים שלמידה לחזות תוצאות השרשה בנוכחות מידע קליני נוסף של גיל הביציות מגבירה את דיוק חיזוי ההשתלה, אך מביאה לדיוק נחות בסווג עוברים בעלי מראה נמוך לגבוה באותה העת. בנוסף, אנו מספקים ניתוח מפורט של יחסי הגומלין בין הרעש בתוויות ההשרשה, מורכבות הדוגמאות וכמויות נתונים בכדי לייעל את ביצועי החיזוי לשתי המשימות. תוצאת המחקר שלנו מצביעה גם על אילו דוגמאות יש לשקול בתהליך הלמידה, מעבר לקבוצת הלימוד המסורתית של עוברים אשר הועברו לרחם. יתר על כן, אנו מציעים מתווה ללמידת מודלים של דירוג עובר המבוססים על מורפולוגיה בלבד, מורפוקינטיקה בלבד ושילוב של שניהם. על ידי השוואת מסגרות הלמידה השונות, אנו מספקים הבנה טובה יותר של נגישות המידע העוברי לאורך זמן.

ממצאינו הופנו ליצירת מערכת דירוג שימושית לזמן אמת במרפאות ובמעבדות פוריות. אנו מציגים פונקציית מטרה אדפטיבית חדשה הניתנת ליישום על נתונים מתוייגים, ללא קשר לאורך הסרטון. עוד מוצעת סכמת לימוד עם תוויות פאסודו מנוגדות המבוססת על מנימזציה של יחסים בין זוגות של עוברים ובכך לכלול בנתוני הלימידה עוברים שלא

הועברו (למשל מושמדים או קפואים), המהווים 85% מהעוברים. לבסוף, אנו מציעים שיטה המשלבת את מסגרות הלמידה המבוססות על מורפולוגיה ומורפוקינטית למסגרת לימוד משולבת המשפרת את שתי המסגרות האישיות.

מודלים אלה אינם מוגבלים לשום זמן או שלב התפתחותי, וניתן ליישם אותם אפילו ביום הראשון להפריה. למיטב ידיעתנו, זהו המחקר הראשון שניצל את כל זרם ה TLI, -ולא רק חלקים קטנים ממנו. אנו מראים, לראשונה, כי הדמיה של הזמן עולה על ציון הדו מימד המסורתי אפילו לעוברים בני 5 ימים. המודלים שלנו עלו על שיטות קודמות, כולל מערכת KIDSCORE-D3 המסחרית, וקבוצה של שמונה אנשי מקצוע בכירים, בניבוי תוצאות השרשה, ניתוח מסלול התפתחותי ובדיקות גנטיות.

לסיכום, המודלים שלנו מאפשרים העברת עוברים מוקדם יותר מ -5 ימים, ובכך מצמצמים את הקשיים הכרוכים בתרבית ממושכת, אשר מקטינה את עומס העבודה וחוסכת משאבים רפואיים. חשוב מכך, זיהוי העוברים הטובים ביותר מאפשר העברות של עוברים בודדים, מבלי להקטין את שיעורי ההשרשה, ובכך להימנע מסיבוכים של הריון מרובי עוברים.

חיזוי טיב התפתחות עובר בעזרת למידת מכונה מתוך תמונות וידאו

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת : איתי ארליך

מנחה: ד"ר אסף זריצקי

הוגש לסנט האוניברסיטה העברית בירושלים 08/2021