Deep Neural Networks Reliably Assess Human Blastocyst Quality and Assist in Predicting Implantation Success upon In-Vitro Fertilization

Pegah Khosravi[1,2], **Ehsan Kazemi**[3], **Qiansheng Zhan**[4], **Marco Toschi**[4], **Jonas E. Malmsten**[4], **Cristina Hickman**[5], **Marcos Meseguer**[6], **Zev Rosenwaks**[4], **Olivier Elemento**[1,2,7], **Nikica Zaninovic**[4*], **and Iman Hajirasouliha**[1,2*]

[1]Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, NY, USA
[2]Caryl and Israel Englander Institute for Precision Medicine, The Meyer Cancer Center, Weill Cornell Medicine, NY, USA
[3]Yale Institute for Network Science, Yale University, CT, USA
[4]The Ronald O. Perelman and Claudia Cohen Center for Reproductive Medicine, Weill Cornell Medicine, NY, USA
[5]Institute of Reproduction and Developmental Biology, Imperial College, Hammersmith Campus, London, UK
[6]Instituto Valenciano de Infertilidad, Universidad de Valencia, Valencia, Spain
[7]WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, NY, USA
*Co-corresponding authors.

ABSTRACT

Conventional human blastocyst selection based on morphological classification is the standard method for assessing embryo quality upon in-vitro fertilization (IVF). The selection process is highly subjective and prone to bias of human judgment and differences in perception among embryologists. Although this method is widely used in clinical practice, it does not necessarily provide an accurate estimation of embryo implantation and live-birth potential.
We postulated that an unbiased artificial intelligence (AI) approach trained on a large number of embryos (i.e. thousands) together with known associated clinical data can reliably predict embryo quality without human intervention.
Our AI approach is based on deep neural networks (DNN). We present a computational framework called STORK to accurately predict morphological quality for blastocysts based on raw digital images of fertilized embryos and their associated clinical data. The STORK framework achieves an accuracy of 98% for discrimination between good-quality and poor-quality blastocysts as assessed by embryologists, thus indicating that a DNN can automatically and accurately grades embryos based on raw images.
Using clinical data for 2,182 embryos, we then created a decision tree that integrates clinical parameters such as embryo quality and patient age to identify scenarios associated with increased or decreased pregnancy chance. This data-driven analysis shows that pregnancy chance vary from 13.8% (e.g., when the embryo is of poor-quality as assessed by STORK and the age of patient is over 41 years) to 66.3% (e.g., for a good-quality embryo and when the age of patient is less than 37 years). In conclusion, our AI-driven approach provides a novel way to assess embryo quality and uncovers new possible personalized strategies to increase the likelihood of pregnancy using IVF.

Introduction

Infertility remains an unremitting reproductive issue and affects about 186 million people worldwide[1]. Infertility affects approximately 8% of women of child-bearing age in the United States[2]. Approximately 44% of women in the United States meet criteria for infertility at a certain point during their reproductive years[3]. Assisted reproductive technology (ART), including in-vitro fertilization (IVF), is one of the most common fertility treatments for infertility. IVF involve ovarian stimulation and multiple oocytes retrieval from the growing follicles, fertilization and embryo culture for 1 to 6 days in controlled environmental conditions. Embryo quality is then assessed to select one or more embryos for transfer to the patient's uterus. One of the reasons for transferring multiple embryos is the unavailability of highly accurate and reliable method for selecting good-quality embryos[4]. IVF and embryo transfer technologies have considerably improved over the past 30 years. However, the efficacy of IVF remains relatively low and needs to be improved.[5]

Common practice of embryo evaluation involve observation, evaluation and manual grading of morphological features of embryo (blastocyst) by skilled embryologists for the purpose of selection and transfer. While this method is universally used in the clinical practice, the erroneous evaluation of single static image and decision-making using a rough estimation of embryos can be subjective and time-consuming[6–8].

There is also diversity and a lack of consistency among different clinics in classification of blastocyst quality and associated grading systems. This inconsistency has made comparing methods from different clinics and analyzing patients when they undergo treatments in different clinics very challenging. To date attempts to reach a consensus and universal grading and selection system have failed[9].

Improving the ability to determine which embryos have the highest implantation potential would help increase the pregnancy success rates. It will also minimize the chance of multiple births due to transferring multiple embryos as a way to increase success rate[10]. Opportunities exist to leverage AI since IVF clinics have long adopted digital imaging as part of their clinical practice and have accumulated thousands of labeled images and time-lapse datasets.

Time-lapse imaging (TLI) is an emerging technology that allows continuous observation of embryo development without removal from controlled and stable incubator condition[11]. Time-lapse analysis was first used more than three decades ago for the study of the developmental progression of bovine embryos in vitro[12,13]. Recent interest in this technology for assessment of clinical embryos is due to improved selection of the most robust embryos for transfer[14]. This technology also improved outcomes of IVF cycles by decreasing risk of keeping embryos in a disturbed environment such as temperature changes, high oxygen exposures, and pH changes during culture[15]. In addition, it enabled embryologists to assess the quality of embryos by tracking the timing of embryo cleavage events and length of different intervals in embryo development (karyokinesis and cytokinesis)[16].

Currently no robust and fully automatic method exists to analyze human embryo data from TLI. A few groups attempted to use various machine learning approaches for embryo quality analysis, with variable level of success[17,18] for bovine and mammalian oocytes using artificial neural network (ANN) and random forest (RF) based classification, respectively. Their results showed 76.4% (test set = 73 embryos), 75% (test set = 56 embryos) accuracy for discretization of bovine embryo grades (excellent, fair, and poor) and mammalian oocytes grades (A, B, C, and D), respectively. Furthermore, a few previously published approach has focused on classification of human embryo quality based on specific features, such as Inner Cell Mass (ICM) area, Trophectoderm (TE) area, and zona pellucida (ZP) thickness and blastocyst area and radius separately[10,19]. In particular, Filho et. al.[19] presented a semi-automatic grading of human embryos. They showed the classifiers can have different accuracies for each object (blastocyst extension, ICM, and TE). Their results indicated various accuracy ranges from 67% to 92% for the embryo extension, 67% to 82% for the ICM, and 53% to 92% for the TE detection and 92% was the highest accuracy achieved across 73 embryos as test set[19]. Although these methods achieved a reasonable accuracy in assessing human embryo quality, they need enormous embryology experience and a lot of preprocessing steps which is time consuming.

Deep learning has recently been used to address a number of medical imaging problems, such as predicting skin lesions or diagnosing disease[20]. Our group also recently showed that deep learning can significantly improve performance, correctness, and robustness in classification and quality assessment of digital pathology images in cancer[21].

In this paper, we introduce a computational method using deep learning techniques (Figure 1) to predict quality of human embryos. In the first step, our embryologists generated embryo images from time-lapse imaging and manually labeled them to two classes of good-quality or poor-quality. This was performed using reanalysis of existing embryo grades in the context of live-birth information (see Methods). In the second step, a deep neural network is trained to assess the quality of images automatically. Finally, a decision tree is used to combine the deep learning-based assessment of embryo quality with clinical data such as patient's age to identify (ideal) clinical scenarios that are associated with maximized likelihood of pregnancy (Figure 1). We evaluate the performance of our method using a blind test set comprises good- and poor-quality image of human embryos.

Results

We obtained time-lapse images for 10,148 embryos, taken at 110hpi after fertilizing oocytes, at the Center for Reproductive Medicine at Weill Cornell Medicine, New York (WCM-NY). Images were taken at seven focal depths (+45, +30, +15, 0, -15, -30, -45), constituting a set of 50,392 images in total.

Trained embryologists evaluated embryo quality using an internal scoring system with 130 distinct classes. To enable the AI analysis, the 10,148 embryos were subsequently classified into 3 major groups (good-quality = 1,345 embryos, fair-quality = 4,062 embryos, poor-quality = 4,741 embryos) (Figure 2a) as described in Methods.

We sought to train an Inception-V1 deep learning based algorithm using the two quality groups at both ends of the spectrum (good and poor). The Inception-V1 architecture is a transfer learning algorithm and we initially performed fine-tuning of parameters for all the layers. Upon preprocessing and removal of bad quality images and random selection of balanced sets of images, we were left with 12,001 images with up to seven focal depths (+45, +30, +15, 0, -15, -30, -45) of good- (6,000 images, 877 embryos) and poor-quality (6,001 images, 887 embryos) labels. We

used 50,000 steps for training the DNN. We then evaluated the performance of STORK using a randomly selected independent test set with 964 good-quality (141 embryos) and 966 poor-quality embryo (142 embryos) images.

DNN architecture achieves expert-level classification in embryo images

Our results showed that the trained algorithm is able to identify good-quality and poor-quality images with 96.94% accuracy (1,871 correct prediction out of 1,930 images = 96.94% accuracy) when tested on 964 good-quality and 966 poor-quality images.

To measure the accuracy of STORK for embryos with multiple image focal depths, we used a simple voting system. If the majority of images from the same embryo is good then the final quality of the embryo is considered good. For a small number of cases where the number of good and poor images is equal (e.g., 3 good and 3 poor when the number of focal depth is 6) we use STORK's output probability scores for tie-breaking. We compared the average STORK probability scores of the good images with the average probability score of poor images.

We observed 97.53% accuracy (276 correct prediction out of 283 embryos = 97.53% accuracy), (Figure 2b) (comprises 283 embryos) as a blind test set. We also found that training an Inception-V1 model without fine-tuning did not affect performances (accuracy). See Figure Supplementary 1. This observation is in agreement with previous studies using these deep learning techniques[20–22].

We also found that, by using STORK to classify the images with fair-quality (4,480 images from 640 embryos) into one of the two good or poor classes, 82% (526 embryos) and 18% (114 embryos) of the embryos were predicted to be good-quality and poor-quality, respectively (Figure 2c). Attesting to the intermediate status of the fair group, the average probability score was 0.98 for good-quality and 0.93 for poor-quality classes (Figure Supplementary 2), which is significantly (p-value < 0.01) lower than the probability scores for good and poor images (0.99 on average).

Because Inception-V1 was trained for good and poor classes with different successful implantation probabilities (around 58% and 35% pregnancy chances for good and poor classes, respectively), we wondered if STORK nonetheless produced relevant predictions within the fair class. A closer look showed that embryos with fair-quality which are classified to the poor class by STORK had lower likelihood of positive live birth (50.9%) compared to those which are classified to the good class (61.4% positive live-birth, while the statistical significance of the this difference in the outcomes has a p-value < 0.05 by the two-tailed Fisher's test).

We also found that fair embryos predicted to be good-quality by STORK came from younger patients (33.98 years old on average) than predicted poor-quality (34.25 years old on average). Interestingly, these numbers are similar to the good-quality and poor-quality ages which are significantly different (p-value < 0.01) in age 33.86 and 34.72 years old on average, respectively. This suggests that STORK finds sufficient structure within embryos currently classified as fair to make clinical relevant predictions.

Robustness of STORK

In order to evaluate the robustness of STORK, we tested its performance by using additional datasets of embryo images obtained from two other IVF centers, IRDB-IC and Universidad de Valencia, comprising 127 (74 good, 53 poor) and 87 (61 good, 26 poor) embryos, respectively (See Figure 2b). Our experimental results demonstrate that, although the scoring systems used for these centers are different from the system used to train our model, STORK can successfully identify and register score variations and discriminate them robustly with accuracy of 77% (average precision = 0.8, AUC = 0.9) and 70% (average precision = 0.66, AUC = 0.76) for the IRDB-IC and Universidad de Valencia, respectively (Table Supplementary 3).

It is well known that the scoring of embryos frequently differs among embryologists[23]. This happens mainly due to the subjectivity of the scoring process and different interpretations of embryo's quality.

We also applied STORK to additional dataset evaluated by five embryologists from three different clinics. We asked them to provide scores for each of 394 embryos which is generated in different labs. Note that these images are not used in the training phase of our algorithm. The images of embryos were scored using the Gardner scoring system[24], and then mapped onto our simplified three groups (good, fair, and poor) (To find the mapping method see Table Supplementary 5 and Table Supplementary 2).

We found, surprisingly, a low level of agreement among the five embryologists, with only 89 embryos out of 394 embryos labeled with the same quality by all the five embryologists (Figure Supplementary 3). Therefore, to create a larger and more accurate gold-standard dataset, we used an embryologist majority voting procedure (i.e., the label of each image is the label reported by at least three out of five embryologists) to label images consisting of 239 images (32 good and 207 poor).

When we applied STORK to all these 239 images, we found that STORK predicted the embryologist majority vote with high accuracy (90.4%) and average precision (95.7%). In comparison, STORK agreed with the individual embryologists slightly less often (89.6%, 85.8%, 80.8%, 85.8%, 88.3% accuracy, 92.1%, 89.5%, 97.4%, 88.3%, 96.3%

average precision). These results indicate that STORK is at least as reliable as any individual embryologist when classifying embryo image quality (Figure 2d).

Predicting pregnancy likelihood using the trained algorithm for embryos' outcome

It is known that different factors such as the embryo quality, maternal age, patient's genetic background, clinical diagnosis, and treatment related characteristics can affect the pregnancy outcome[25,26].

Because the embryo quality is one of the most important characteristics that can affect the pregnancy rate, the ultimate aim of any embryo quality assessment approach is to identify embryos with the highest implantation potential and resulting in live birth.[24,27,28].

We explored the possibility of predicting the pregnancy likelihood based on embryo morphologic quality using images that are labeled as positive or negative live-birth. In other words, we wondered how much of the pregnancy rate is associated to the morphological quality of embryos.

To address this question, we used WCM-NY images associated to the 1,620 available embryos that have positive and negative pregnancy outcome (live-birth) information (Table Supplementary 1).

We allocated 85% of the embryos (1,377 embryos, 9,639 images) to build two classes containing "negative live-birth" (603 embryos) and "positive live-birth" (774 embryos) as training. Each class contains embryos with good- and poor-quality. It means we have good-quality and poor-quality embryos in "negative live-birth" (embryos 'a' and 'b' at Figure Supplementary 4). Also, "positive live-birth" class comprises embryos with poor-quality and good-quality (embryos 'c' and 'd' at Figure Supplementary 4). Thus, we have embryo images with four different characteristics in two classes (Figure Supplementary 4).

We build a new training algorithm DCNN (deep convolutional neural network), different from STORK, to fine-tune Inception-V1 algorithm using two classes (positive and negative live-birth) with 50,000 steps.

Finally, we tested DCNN with 243 randomly selected embryos as a blind test comprising 136 and 107 embryo (1,701 images) as "positive" and "negative", respectively (Table Supplementary 1).

We obtained only a 51.85% accuracy for discretization of positive and negative live-birth. This suggests that discretization of images based on live-birth outcome using their morphology alone cannot be useful since other important characteristics such as patient's age and genetic or clinical variations can affect the pregnancy rate.

Therefore, in the next section we present an alternative method for pregnancy probability prediction based on a state-of-the-art decision tree method via integration of clinical information and embryo quality.

Decision tree reveals the interaction among clinical information

As we showed in the previous section, the quality of embryos are not enough to accurately determine the pregnancy probability. Fortunately, there are other clinical variations that affect the pregnancy likelihood. Therefore, we wonder if we can asses the pregnancy rate using the combination of embryo quality and patient's age, as one of the most important clinical variables. For this purpose, we used a hierarchical type class of decision trees[29] known as Chi-Squared Automatic Interaction Detection (CHAID) algorithm.

We designed a CHAID[30,31] decision tree using all 2,182 embryos from the WCM-NY database with available clinical information. We then investigated the interaction between patient's age (consist of seven classes of 30 or younger, 31 - 32, 33 - 34, 35 - 36, 37 - 38, 39 - 40, and older than 41) and the embryo quality (two classes of good and poor) on live-birth outcome. We used 1,620 embryos that treated through IVF treatment types. The CHAID algorithm can project interactions between variables, and non-linear effects which are generally missed by traditional statistical techniques. CHAID builds a tree in order to determine how variables can explain the outcome in a statically meaningful way.[30,31]. CHAID uses χ^2 statistics through the identification of optimal multi-way splits, and identifies a set of characteristics (e.g., patient's age and embryo quality) that best differentiates individuals based on a categorical outcome (here is live-birth) and creates exhaustive and mutually exclusive subgroups of individuals. It chooses the best partition on the basis of statistical significance and uses Bonferroni adjusted p-values to determine significance with a predetermined minimum size of end nodes. We used 1% Bonferroni adjusted p-value, maximum depth of the tree (n = 5), and a minimum size of end nodes (n = 20) as the stopping criteria. The application of tree-based algorithm on the embryo data would help to more precisely define the patient's age and embryo quality (good or poor) effect on live-birth outcome and to better understand any interactions between these two clinical variables (patient's age and embryo quality).

Note that, while several other classification algorithms can also be employed, the use of CHAID in the prediction presented the best fit in both model quality criteria and more proper decision tree diagram visually[32,33].

As Figure 3 shows, patients are classified in three groups based on their age: (i) 36 year or younger, (ii) 37 and 38 years old, and (iii) 39 years or older. Embryos in each age group with good- and poor-quality are discretized in two groups.

The result confirms the association between pregnancy probability of patients with their ages. The pregnancy probability of embryos with good-quality is significantly (1% Bonferroni adjusted p-value) higher than patients with poor-quality embryos across different ages. Figure 3 indicates, the patients who are 36 year or younger have a higher pregnancy rate compare to the the pregnancy likelihood for patients in two other age groups.

Discussion

The era of computational embryology is rapidly evolving and there are enormous opportunities for computational approaches to provide additional prognostic information that cannot be provided by embryologists alone. The STORK framework presented here provides a novel method that can be easily implemented for a wide range of applications, including embryo grading.

Recently several papers have been published that utilize various methods such as classical machine learning approaches including support vector machine (SVM) and random forest, and deep learning methods such as CNN-basic[17,18,34] for prediction of outcome or grade classification. Several artificial intelligence (AI) methods have been used to assess blastocysts to date[35]. Image segmentation and advanced image analysis techniques using neural networks with textured descriptors, level set, phase congruency, fitting of ellipse methods, have been demonstrated in mice[36], bovine[17], and human blastocysts[19,37]. Studies on human embryos are yet very limited. They often involve low numbers (51 to 394) of embryos from single centers and lack desired validations. Furthermore, the publications to date relied on images that were captured using inverted microscopes. However, Time-Lapse images have the advantage of being consistent both in terms of image size, lighting, contrast and quality; and in terms of timing of embryo development, particularly important when quantifying blastocyst expansion which is particularly time-dynamic.

The aim of this project is to evaluate the utility of deep neural networks to automatically identify embryo quality. To the best of our knowledge, this is the first report that utilize higher level architecture of a DNN algorithm and compare its performance on embryo images across various configurations. The advantage of this technique is that the whole image of the embryo is assessed, not only pre-determined features segmented, allowing for quantification of all the data available, not just the part we are trained to focus on based on prior knowledge bias. Convolution, therefore, allows the AI to identify patterns in morphological features we did not know how to assess.

We indicated that deep learning approaches can provide accurate quality assessments in various clinical conditions. Our results show the accuracy of a deep neural network primarily depends on the selected labels that we train the algorithm with them. We also defined a gold standard classification system to map the quantitative number of grades from 130 different grades to two (good, poor) quality grades.

Our method yields cutting edge sensitivity on the challenging task of detecting various embryo classes in embryo slides, reducing the false rate. Note that, our STORK framework requires no prior knowledge of an image color space or any parameterizations from the users and is fully automated. It provides embryologists or medical technicians a straightforward platform to use without requiring sophisticated computational knowledge.

Finally, we designed a decision tree using the CHAID algorithm to investigate interaction between embryo quality and patient's age on the pregnancy rate (live-birth likelihood). This approach can be applied on other clinically relevant parameters influencing IVF outcome.

We also showed that our study raise several important issues regarding embryo conditions since different response of deep learning could be due to clinical situations. Further studies required in order to clarify the efficiency of the deep learning application in detection of pregnancy outcome.

Methods

In this section we first present our AI-based method for classifying embryo morphologies. We also discuss how we assessed the accuracy and consistency of the AI classifier in comparison to human classification.

Embryo images resource

This study included 10,148 embryos from our Center for Reproductive Medicine at Weill Cornell Medicine (2012/05-2017/12). We refer to this dataset as WCM-NY throughout this manuscript. The images were captured using the following technique: EmbryoScope® time-lapse system (Vitrolife, Sweden) Built-in microscope: Leica 20x, 0.40 LWD Hoffman modulation contrast objective specialized for 635 nm illumination, Camera resolution: 1280×1024 pixels, three pixels per µm, monochrome, 8-bit. Embryo illumination: 0.032s per image using single red LED (635nm) gives 34µW cm-2 for image acquisition. Time between acquisitions: 15 min. cycle time for seven focal planes representing a total of 50,392 images (stored in jpg, 500×500 pixels) with about seven focal depths (+45, +30, +15, 0, -15, -30,

-45) that are captured precisely 110 hours post-insemination (hpi) (Figure Supplementary 5). The standardization of images by the Embryoscope software was consistent and images are labeled using Veeck and Zaninovic grading systems[38]. Also, these images contain 130 various grades that most of them comprise a few image number for each grade (Table Supplementary 4). We eliminated images that are either very dark or missing embryo picture from the dataset, and selected a balanced set of images for both good and poor classes.

The Veeck and Zaninovic grading systems[38] (Table Supplementary 5) is a slightly modified classification system of the Gardner system[24], classifying the embryos based on blastocyst expansion (grades 1 to 6), cell abundance and conformity in the Inner Cell Mass (grades A, B, C) and Trophectoderm (grades A, B, C) (Table Supplementary 5).

In addition to our WCM-NY data, we used two other data sets from the Universidad de Valencia and the Institute of Reproduction and Developmental Biology of Imperial College (IRDB-IC). The data from the Universidad de Valencia was graded based on a slightly different scoring system known as Asebir[39]. Compared to the Gardner system, Asebir reduces the expansion categories to five categories (rather than six) and changes the ICM and trophectoderm rating terminology to A, B, C, and D letters (Table Supplementary 5). The IRDB-IC data was graded using the Gardner scoring system.

Classification and diagnostic framework

This study presents a framework (see Figure 1) to classify different embryo images based on Veeck and Zaninovic grades (Table Supplementary 4) and map the grades to good and poor quality blastocyst grades. Here we used the WCM-NY embryos and the clinical information of a subset of the embryos such as grades and patients' age .

We partitioned the images into training, validation, and test groups. We allocated 70% of the images to the training group, while the remaining 30% of the images were devoted to validation and test sets. The training, validation, and test sets do not overlap.

Algorithm architectures and training methods

We employed a deep neural network (DNN) for embryo image analysis based on Google's Inception-V1[40] architecture, which offers a very effective run-time and computational cost[41,42]. To train this architecture, we used transfer learning. We employed a pre-trained network and fine tuned all outer layers[43] using WCM-NY images. We also compared this transfer learning approach to training the network from scratch.

Evaluation of method and Implementation details

To implement the STORK framework, we used Tensorflow version: 1.4.0[44] and the Python library TF-Slim for defining, training, and evaluating models in TensorFlow. All training of our deep learning methods were performed on a server running SMP Linux operating system. This server is powered by four NVIDIA GeForce GTX 1080 GPUS with 8 GB of memory for each GPU and twelve 1.7 GHz Intel Xeon CPUs.

To evaluate the performance of our methods, we used an *accuracy* measure which is the fraction of correctly identified images[21]. The accuracy is formally defined as TNu /(TNu + FNu), where TNu (true number) and FNu (false number) are the number of correctly and incorrectly classified images.

To assess the performance of different algorithms precision-recall curves (PRCs) are used. Here, precisions and recalls are presented by average for multi-class datasets. Additionally, receiver operating characteristics (ROCs) were estimated. The ROC curve depicts by plotting the true positive rate (TPR) versus the false positive rate (FPR) at various threshold settings. The accuracy is measured by the area under the ROC curve (AUC)[45,46].

Code availability

The trained algorithm (STORK), source code, training manual steps, and the training data sets are publicly available at `http://github.com/ih-lab/STORK`.

References

1. Inhorn, M. C. & Patrizio, P. Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century. *Hum. reproduction update* **21**, 411–426 (2015).
2. Chandra, A., Copen, C. E. & Stephen, E. H. *Infertility and impaired fecundity in the United States, 1982-2010: data from the National Survey of Family Growth.* 2013 (US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2013).
3. McQuillan, J., Greil, A. L. & Shreffler, K. M. Pregnancy intentions among women who do not try: Focusing o n women who are okay either way. *Matern. child health journal* **15**, 178–187 (2011).

4. Ajduk, A. & Zernicka-Goetz, M. Advances in embryo selection methods. *F1000 biology reports* **4** (2012).
5. Dyer, S. *et al.* International Committee for Monitoring Assisted Reproductive Technologies world report: assisted reproductive technology 2008, 2009 and 2010. *Hum. reproduction* **31**, 1588–1609 (2016).
6. Conaghan, J. *et al.* Improving embryo selection using a computer-automated time-lapse image analysis test plus day 3 morphology: results from a prospective multicenter trial. *Fertility sterility* **100**, 412–419 (2013).
7. Paternot, G., Debrock, S., De Neubourg, D., d'Hooghe, T. & Spiessens, C. Semi-automated morphometric analysis of human embryos can reveal correlations between total embryo volume and clinical pregnancy. *Hum. reproduction* **28**, 627–633 (2013).
8. Tian, Y. *et al.* Predicting pregnancy rate following multiple embryo transfers using algorithms developed through static image analysis. *Reproductive biomedicine online* **34**, 473–479 (2017).
9. Puga-Torres, T., Blum-Rojas, X. & Blum-Narváez, M. Blastocyst classification systems used in latin america: is a consensus possible? *JBRA assisted reproduction* **21**, 222 (2017).
10. Saeedi, P., Yee, D., Au, J. & Havelock, J. Automatic identification of human blastocyst components via texture. *IEEE Transactions on Biomed. Eng.* **64**, 2968–2978 (2017).
11. Chen, M., Wei, S., Hu, J., Yuan, J. & Liu, F. Does time-lapse imaging have favorable results for embryo incubation and selection compared with conventional methods in clinical in vitro fertilization? a meta-analysis and systematic review of randomized controlled trials. *PloS one* **12**, e0178720 (2017).
12. Massip, A. & Mulnard, J. Time-lapse cinematographic analysis of hatching of normal and frozen—thawed cow blastocysts. *J. Reproduction Fertility* **58**, 475–478 (1980).
13. Massip, A., Mulnard, J., Vanderzwalmen, P., Hanzen, C. & Ectors, F. The behaviour of cow blastocyst in vitro: cinematographic and morphometric analysis. *J. anatomy* **134**, 399 (1982).
14. Finn, A., Scott, L., O'Leary, T., Davies, D. & Hill, J. Sequential embryo scoring as a predictor of aneuploidy in poor-prognosis patients. *Reproductive biomedicine online* **21**, 381–390 (2010).
15. Racowsky, C., Kovacs, P. & Martins, W. P. A critical appraisal of time-lapse imaging for embryo selection: where are we and where do we need to go? *J. assisted reproduction genetics* **32**, 1025–1030 (2015).
16. Armstrong, S., Vail, A., Mastenbroek, S., Jordan, V. & Farquhar, C. Time-lapse in the ivf-lab: how should we assess potential benefit? *Hum. Reproduction* **30**, 3–8 (2014).
17. Rocha, J. C. *et al.* A method based on artificial intelligence to fully automatize the evaluation of bovine blastocyst images. *Sci. reports* **7**, 7659 (2017).
18. Viswanath, P., Weiser, T., Chintala, P., Mandal, S. & Dutta, R. Grading of mammalian cumulus oocyte complexes using machine learning for in vitro embryo culture. In *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*, 172–175 (IEEE, 2016).
19. Filho, E. S. *et al.* A method for semi-automatic grading of human blastocyst microscope images. *Hum. Reproduction* **27**, 2641–2648 (2012).
20. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nat.* **542**, 115 (2017).
21. Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. & Hajirasouliha, I. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine* (2017).
22. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* **316**, 2402–2410 (2016).
23. Arce, J.-C. *et al.* Interobserver agreement and intraobserver reproducibility of embryo quality assessments. *Hum. Reproduction* **21**, 2141–2148 (2006).
24. 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 sterility* **73**, 1155–1158 (2000).
25. Subira, J. *et al.* Grade of the inner cell mass, but not trophectoderm, predicts live birth in fresh blastocyst single transfers. *Hum. Fertility* **19**, 254–261 (2016).
26. Irani, M. *et al.* Morphologic grading of euploid blastocysts influences implantation and ongoing pregnancy rates. *Fertility sterility* **107**, 664–670 (2017).

27. Kinzer, D. R., Barrett, C. B., Penzias, A. S., Alper, M. M. & Sakkas, D. Evaluation of a high implantation potential (hip) embryo grading system designed to reduce multiple pregnancy. *J. Reproductive Heal. Medicine* **2**, 11–16 (2016).

28. Yang, Z. *et al.* Selection of single blastocysts for fresh transfer via standard morphology assessment alone and with array cgh for good prognosis ivf patients: results from a randomized pilot study. *Mol. cytogenetics* **5**, 24 (2012).

29. Song, Y.-Y. & Ying, L. Decision tree methods: applications for classification and prediction. *Shanghai archives psychiatry* **27**, 130 (2015).

30. Kass, G. V. An exploratory technique for investigating large quantities of categorical data. *Appl. statistics* 119–127 (1980).

31. Hébert, M., Collin-Vézina, D., Daigneault, I., Parent, N. & Tremblay, C. Factors linked to outcomes in sexually abused girls: a regression tree analysis. *Compr. psychiatry* **47**, 443–455 (2006).

32. Ali, M. *et al.* Comparison of artificial neural network and decision tree algorithms used for predicting live weight at post weaning period from some biometrical characteristics in harnai sheep. *Pak. J. Zool.* **47** (2015).

33. Chen, W. *et al.* Establishing decision trees for predicting successful postpyloric nasoenteric tube placement in critically ill patients. *J. Parenter. Enter. Nutr.* **42**, 132–138 (2018).

34. Jeanray, N. *et al.* Phenotype classification of zebrafish embryos by supervised learning. *PloS one* **10**, e0116989 (2015).

35. Santos Filho, E., Noble, J. & Wells, D. A review on automatic analysis of human embryo microscope images. *The open biomedical engineering journal* **4**, 170 (2010).

36. Matos, F. D., Rocha, J. C. & Nogueira, M. F. G. A method using artificial neural networks to morphologically assess mouse blastocyst quality. *J. animal science technology* **56**, 15 (2014).

37. Manna, C., Nanni, L., Lumini, A. & Pappalardo, S. Artificial intelligence techniques for embryo and oocyte classification. *Reproductive biomedicine online* **26**, 42–49 (2013).

38. Veeck, L. L. & Zaninovic, N. *An atlas of human blastocysts* (Taylor & Francis, 2003).

39. Saiz, I. C. *et al.* The embryology interest group: updating asebir's morphological scoring system for early embryos, morulae and blastocysts. *Medicina Reproductiva y Embriologia Clin.* **5**, 42–54 (2018).

40. Szegedy, C. *et al.* Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9 (2015).

41. Movshovitz-Attias, Y. *et al.* Ontological supervision for fine grained classification of street view storefronts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1693–1702 (2015).

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

43. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Analysis* **42**, 60 – 88 (2017). DOI https://doi.org/10.1016/j.media.2017.07.005.

44. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. In *OSDI*, vol. 16, 265–283 (2016).

45. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiol.* **143**, 29–36 (1982).

46. Zawistowski, M. *et al.* Corrected roc analysis for misclassified binary outcomes. *Stat. Medicine* **36**, 2148–2160 (2017).

Acknowledgements

We acknowledge Dr. Fabien Campagne for useful discussions and providing additional computing resources for our analysis. This work was supported by start-up funds (Weill Cornell Medicine) to IH. EK was supported by Swiss National Science Foundation under grant number 168574.

Author contributions statement

PK, EK, JEM, CH, MM, NZ, OE, and IH conceived the study. PK, EK, OE and IH conceived the method and designed the algorithmic techniques. QZ, MT, CH, MM and NZ generated the data sets, prepared and labeled the images for various grades. PK and EK wrote the codes and performed computational analysis with input from OE and IH. ZR provided critical reading and suggestion. PK, EK, QZ, OE, NZ, and IH wrote the paper and all authors read, edited and approved the final manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Figures

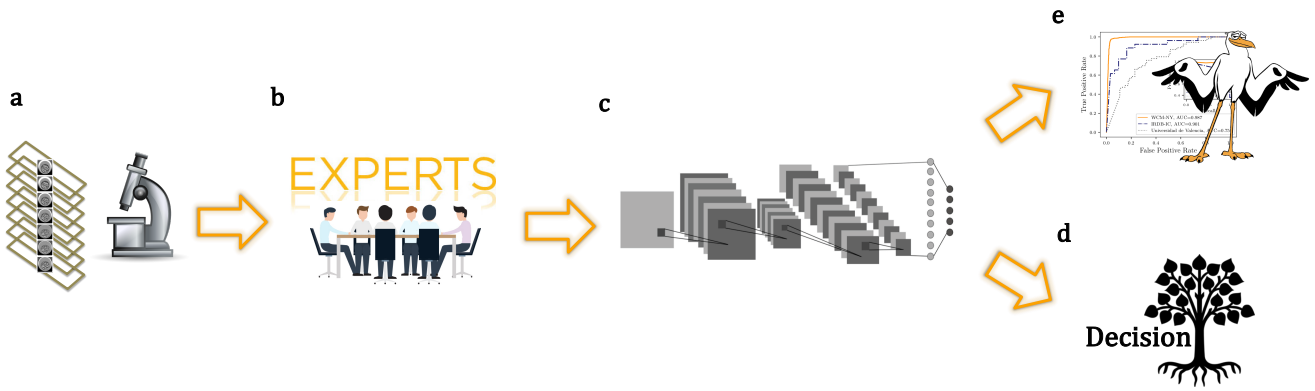
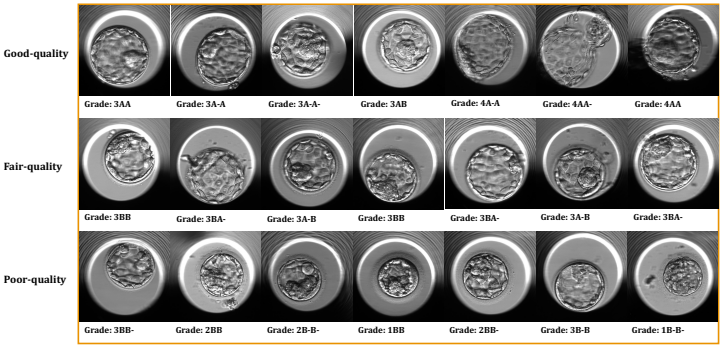
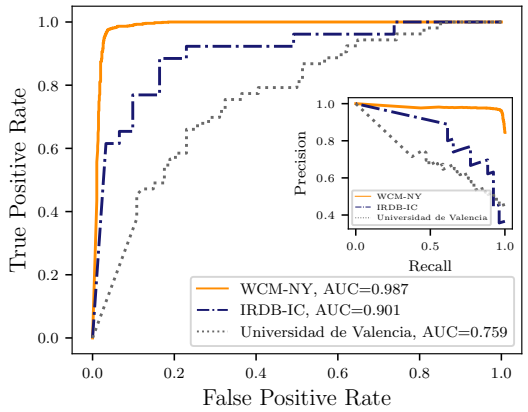


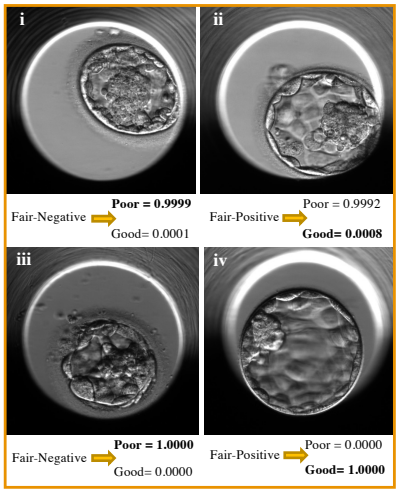
Figure 1. This flowchart demonstrates the design and assessment of STORK (a) providing human embryo images from the embryology lab, (b) embryo images are labeled as good or poor based on their pregnancy likelihood by embryologists, (c) the labels and clinical information from the extracted images are integrated and the Inception-V1 algorithm is trained for good and poor classes, (d) the CHAID decision tree is used to investigate the interaction between clinical information such as patient's age with embryos' quality, (e) STORK is evaluated by a blind test set to assess its performance for prediction of embryos' quality.



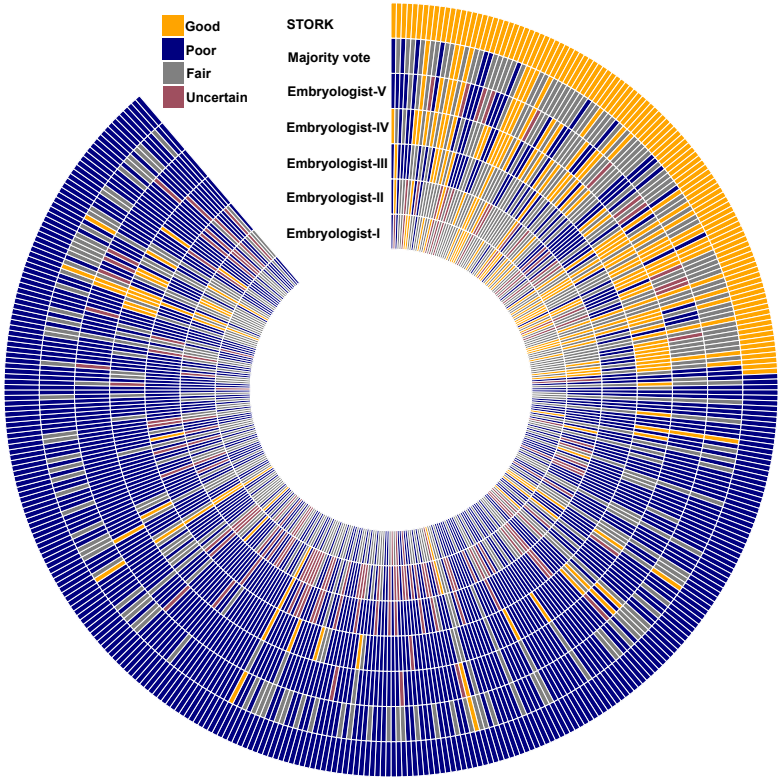
(a) Different morphological quality of embryos



(b) Inception-V1 (fine-tuning)



(c) Classifying images with fair-quality



(d) Agreement between STORK and five embryologists

Figure 2. (a) Embryologists evaluate embryo quality using an internal scoring system and subsequently classify them into three major groups (good-quality, fair-quality, poor-quality). (b) Inception-V1 (fine-tuning the parameters for all layers) results for three datasets. WCM-NY: Shows the data from the center for Reproductive Medicine and Infertility at Weill Cornell Medicine of New York; Universidad de Valencia: Shows the data from the Institute Valenciano de Infertilidad, Universidad de Valencia; IRDB-IC: Shows the data from the Institute of Reproduction and Developmental Biology of Imperial College. (c) STORK classifies the images with fair-quality into existing poor and good classes. For example, figures 'i' and 'ii' are labeled to 3A-B of Veeck and Zaninovic, while STORK classified them to poor and good, respectively. Also, figures 'iii' and 'iv' are both labeled to 3BB. However, the algorithm correctly classified figure 'iii' as poor and figure 'iv' as good. As the figure shows, 'ii' and 'iv' embryo's outcome is positive live-birth while 'i' and 'iii' embryo's outcome is negative live-birth. (d) The circular heatmap demonstrates the agreement between STORK and five embryologists in labeling of the same images form 394 embryos. Also, the heatmap compares STORK's result with the majority vote results from all the embryologists on 239 embryos. The orange color indicates the embryos with good-quality, the navy color shows the embryos with poor-quality, and the red color demonstrates the embryos with fair-quality. Also, the gray color shows the embryos which are not labeled due to uncertainty.

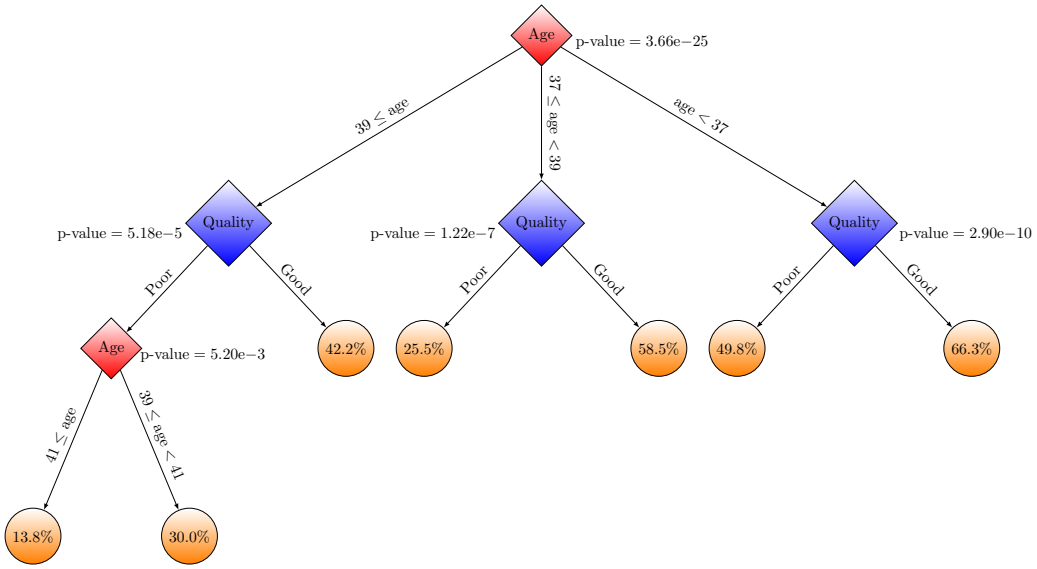


Figure 3. The decision tree shows the interaction between IVF patient's ages and embryos' quality using CHAID.

Supplementary Information

Embryologists splits and merges quantity grades

In this project, the quantitative scores are determined by skilled embryologists based on the grading system of Veeck and Zaninovic[38]. This grading system has three components: The first component is a number showing the level of blastocyst expansion (CM, 1, 2, 3, 4, 5), the second component is a letter indicating the cell abundance and conformity in the ICM (grades A, B, C, D) and finally the third component is a letter quantifying the quality of TE cells (grades A, B, C, D) that are extra-embryonic tissues with supporting role for the embryo proper (see Table Supplementary 5). For the first step of this project, the embryologists select 13,931 images based on their pregnancy outcome as embryos with good-quality and poor-quality. The embryologists labeled embryo images to map certain quantitative scores from the grading system of Veeck and Zaninovic (e.g 1BB vs. 3AA) to narrowed quality grades such as just *poor and* good (Table Supplementary 5). In this regard, any score which contains B- or C and the extension rate equal or less than three are considered as a poor group (less than 35% pregnancy chance). Also, any score with two A or A-, or one A with B with extension 3 or greater than 3 can be labeled as good (more than 58% pregnancy chance). However, there are still some scores (e.g., 3BB, 3BA-) that embryology experts debated about them to put them in a separate category (fair-quality) or classify them as good quality since their pregnancy chance are about 48% to 50%. The complete list of scores and their quality map are indicated in Table Supplementary 2. In total, 86 out of 130 scores have images with clinical information and 84 scores contain a small number of images in their cohorts (Table Supplementary 4).

We converted various quantitative grades related to other data resources to the Veeck and Zaninovic[38] scoring system before testing our trained algorithm with other clinical resources (Table Supplementary 3). For instance, the 3AA grade in our WCM-NY dataset is equal to BEaa for the Universidad de Valencia dataset and 4AA for the IRDB-IC dataset which is based on Gardner system[9,24]. Note that, the lower accuracy of these two datasets compared to WCM-NY is due to variations in grading system. The information about the grading systems used for different data sets are given in (Table Supplementary 5).

Predicting pregnancy rate based on morphological quality of embryos

We wondered why the accuracy of DCNN in predicting pregnancy rate via positive and negative live-birth is low. To find the reason, we looked closer at results for embryos with four different characteristics (Figure Supplementary 4) that we integrated into two classes (positive and negative live-birth).

We found that 28.85%, 47.27%, 41.02%, and 71.13% accuracy for a randomly selected test set (243 embryos) comprises "negative live-birth" with "good-quality"(52 embryos) (embryo 'a' at Figure Supplementary 4), "negative live-birth" with "poor-quality" (55 embryos) (embryo 'b' at Figure Supplementary 4), "positive live-birth" with "poor-quality"(39 embryos) (embryo 'c' at Figure Supplementary 4), and "positive live-birth" with "good-quality" (97 embryos) (embryo 'd' at Figure Supplementary 4), respectively.

This suggests the trained algorithm can classify images just based on their quality (good or poor) disregard their outcome (positive or negative live-birth) (Figure Supplementary 4). Therefore, the accuracy of DCNN could be increased if we utilized higher number of images with "poor-quality and negative live-birth" and "good-quality and positive live-birth" in our test set. Moreover, the DCNN performance decreased due to integration of good- and poor-quality images with, for example, "negative live-birth" in a single class (e.g. embryos 'a' and 'b' at Figure Supplementary 4).

Table Supplementary 1

Table Supplementary 1. Four datasets that show different images (different number of embryos and clinical information) are selected from our WCM-NY (three of them) database and the Universidad de Valencia (one of them) database to assess the performance of STORK across different conditions.

Datasets	Dataset representation	Labels of inputs and outputs	Number of classes and images
Good-Poor	110hpi images of embryos (WCM-NY)	Discrimination of good- and poor-quality of embryos	2 classes: 12,001 images for training and 1,930 images for test set
Outcome-Quality	110hpi images of embryos (WCM-NY)	Discrimination of positive and negative outcome of embryos through good- and poor-quality of embryos	2 classes: 9,639 images for training and 1,701 images for test set
Five-Experts	110hpi images of embryos (WCM-NY and Universidad de Valencia)	Discrimination of good- and poor-quality of embryos	2 classes: 12,001 images for training (STORK as trained algorithm by WCM-NY dataset) and 394 embryos for test set from Universidad de Valencia database

Table Supplementary 2

Table Supplementary 2. The table shows the quantity scores that the algorithm is trained for them. The embryologists categorized the scores in two groups (classes) and labeled them as good-quality and poor-quality.

The list of grades	The quality map
3-4AA, 4A-A, 4A-A-, 5AA-, 4AB, 5A-A-, 4AA-, 4AA, 3A-A, 3AA, 3AA-, 3AB, 3A-A-	good-quality
1-2B-/CB, 1-2B-/CB-/C, 1-2B-C, 1B-/CB-/C, 1BC, 1CB-/C, 1CC, 2-2B-C, 2-3BC, 3B-B-/B, 3BC, 3CA-, 3CB, 3CB-, 3CC, 1-2B-/CB-, 1B-/CB, 1B-/CB-, 1B-/CC, 2-3B-/CB, 2-3B-/CB-, 2B-/CB, 3B-/CB-/C, 3B-/CC, 1-2BB-/C, 2B-/CB-/C, 3B-C, 1BB-/C, 1BB-/C, 1-2B-B-/C, 1B-C, 2-3BB-/C, 2-3B-B-/C, 2B-/CB-, 3B-/CB-, 3B-/CB, 2BB-/C, 1B-B-/C, 2B-B-/C, 3BB-/C, 3B-B-/C, 1-2B-B, 1-2B-B-, 2-3B-B-, 1-2BB-, 2-3B-B, 1B-B, 1BB-, 2-3BB-, 1-2BB, 1B-B-, 2B-B, 2B-B-, 2BB-, 3B-B-, 1BB, 2BB, 3B-B, 3BB-	poor-quality

Table Supplementary 3

Table Supplementary 3. The results of applying STORK on various datasets to discriminate two classes of embryo's quality. WCM-NY: The center for Reproductive Medicine and Infertility at Weill Cornell Medicine of New York; Universidad de Valencia: Institute Valenciano de Infertilidad, Universidad de Valencia; IRDB-IC: Institute of Reproduction and Developmental Biology of Imperial College.

Datasets	Grades	Number of test embryos	STORK accuracy
WCM-NY	3AA, 3AA-, 3AB, 3A-A-, 5AA-, 3A-A, 3-4AA, 4AA (good), 3BB-, 2BB-, 2BB, 1BB, 1B-B, 2B-B-, 1BB-, 1B-B-, 2B-B, 1-2BB, 3B-B, 1-2B-B-, 1-2BB-, 1BB-C, 1-2B-B, 3CB (poor)	283	97.53%
Universidad de Valencia	BEab, BEaa, BHiaa, BHiab, BHab (good), BCbb, BCbc, BEcc, BEbc, BCcc, BEcb, BCcb (poor)	127	70.08%
IRDB-IC	4Aa, 4Ab, 5Ab, 5Aa (good), 2Cb, 4Bc, 2Bc, 4Cc, 3Bc, 2Cc, 1Bb, 3Cc (poor)	87	77.01%

Table Supplementary 4

Table Supplementary 4. Characteristics of 130 various grades and their image numbers.

The morphological grades	Class size
1-2B-/CB, 1-2B-/CB-/C, 1-2B-C, 1-2BA-, 1B-/CB-/C, 1BC, 1CB-/C, 1CC, 2-2B-C, 2-3B-A-, 2-3BA, 2-3BC, 2AA, 2AB-/C, 3-4AA, 3A-B-/C, 3B-/CA, 3B-B-/B, 3BC, 3CA-, 3CB, 3CB-, 3CC, 4A-A, 4A-A-, 4AB-, 4B-/CB, 4B-/CB-, 5A-B, 5AA-, 5BA, 5BA-, 5BB-, 1-2A-B-, 1-2B-/CB-, 1B-/CB, 1B-/CB-, 1B-/CC, 2-3AA, 2-3AA-, 2-3B-/CB, 2-3B-/CB-, 2A-B-/C, 2B-/CB, 2BA, 3A-C, 3B-/CB-/C, 3B-/CC, 4B-B-, 4BB-, 5B-B, 1-2BB-/C, 1AB, 2B-/CB-/C, 3B-C, 4AB, 4B-B, 5A-A-, 5BB, 6BB, 1BB-/C, 1BB-/C, 1-2B-B-/C, 1A-B-, 1B-C, 2-3AB-, 2-3BB-/C, 4A-B, 4AA-, 4BA-, 2-3B-B-/C, 2A-A-, 2B-/CB-, 3B-A, 4AA, 2-3BA-, 1-2A-B, 1A-B, 2BA-, 3B-/CB-, 2AB, 5B-B-, 2AB-, MOR	Less than 10 images per grade
2-3A-B-, 3B-/CB, 2A-B-, 2BB-/C, 3B-A-, 4BB, 2-3AB, 2-3A-A-, CM, CAVM, 1B-B-/C, 2B-B-/C, 3BB-/C, 3BA, 3B-B-/C, 2A-B, 1-2B-B	More than 10 and less than 50 images per grade
2-3A-B, 1-2B-B-, 2-3B-B-, 3A-A, 1-2BB-, 3AB-, 2-3B-B	More than 50 and less than 100 images per grade
1B-B, 1BB-, 2-3BB-, 1-2BB, 1B-B-, 3A-B-, 3AA, 2B-B, 3BA-, 3AA-, 2B-B-, 2-3BB, 3AB, 2BB-, 3B-B-, 1BB, 3A-A-	More than 100 and less than 500 images per grade
2BB, 3B-B, 3BB-, 3A-B, 3BB	More than 500 images per grade

Table Supplementary 5

Table Supplementary 5. Information about different grading systems in three clinics.

Objects	Veeck and Zaninovic	Gardner	Asebir
Expansion			
	CM	1	BT
	1	2	BT
	2	3	BC
	3	4	BE
	4	5	BHi
	5	6	BH
ICM			
	A	A	A
	B	A/B	B
	C	B/C	C
	D	C	D
TE			
	A	A	A
	B	A/B	B
	C	B/C	C
	D	C	D

Figure Supplementary 1

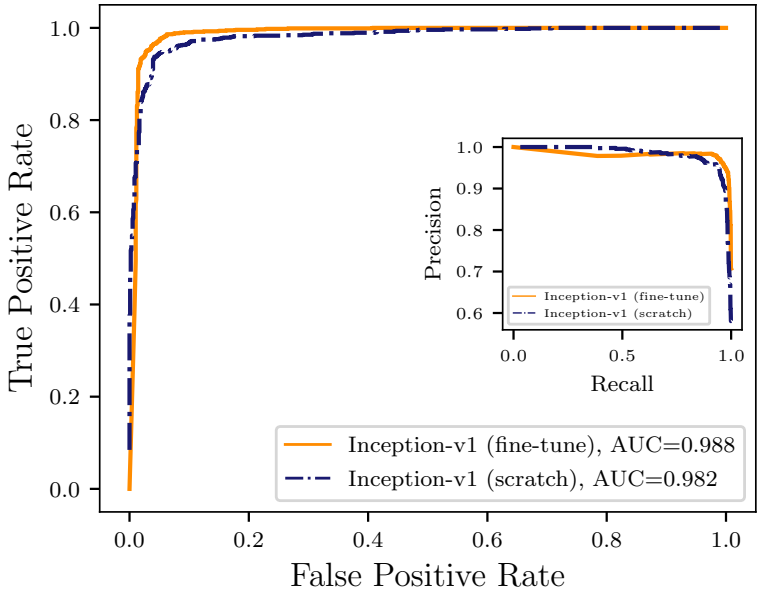


Figure Supplementary 1. Inception-V1 via two different training methods (fine-tuning the parameters for all layers, and training from the scratch) in good and poor embryo quality discrimination dataset.

Figure Supplementary 2

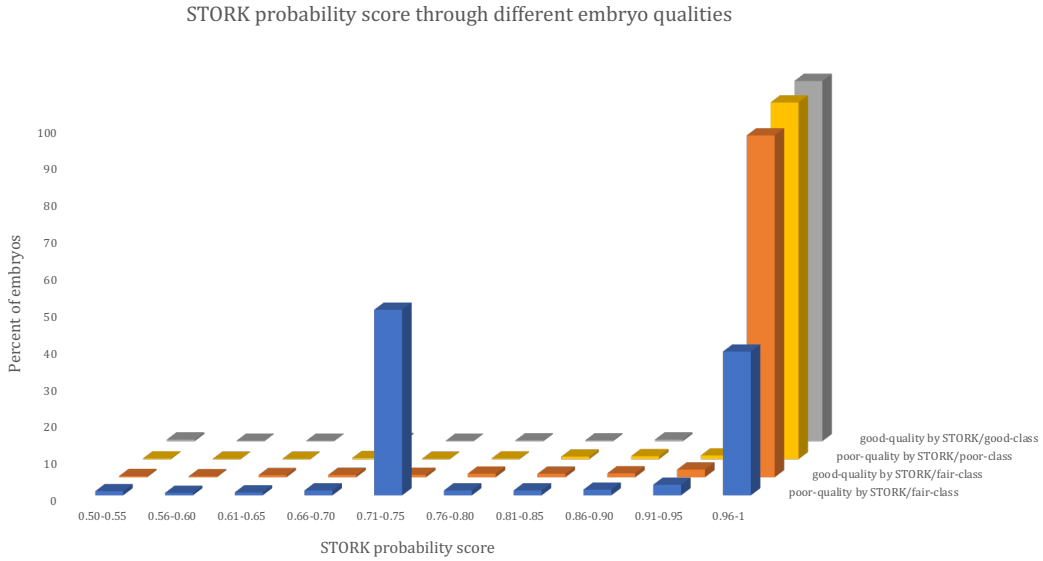


Figure Supplementary 2. STORK gives each embryo of fair class a probability score and classify them into two groups as good-quality and poor-quality. While the score of those embryos that are relabeled by STORK as good and poor is 0.98 for good-quality and 0.93 for poor-quality, the average probability score for both good and poor classes that are labeled by embryologists as good-quality and poor-quality is 0.99.

Figure Supplementary 3

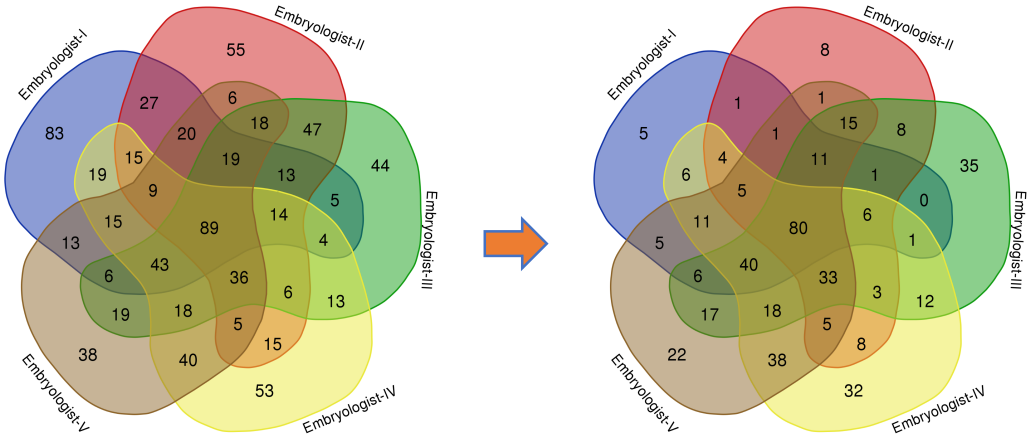


Figure Supplementary 3. This diagram demonstrates the agreement among embryologists (Venn diagram in the left side) and agreement between STORK and five embryologists (Venn diagram in the right side) in labeling of the same embryo images. Different colors indicate different embryologists and numbers show the number of embryos.

Figure Supplementary 4

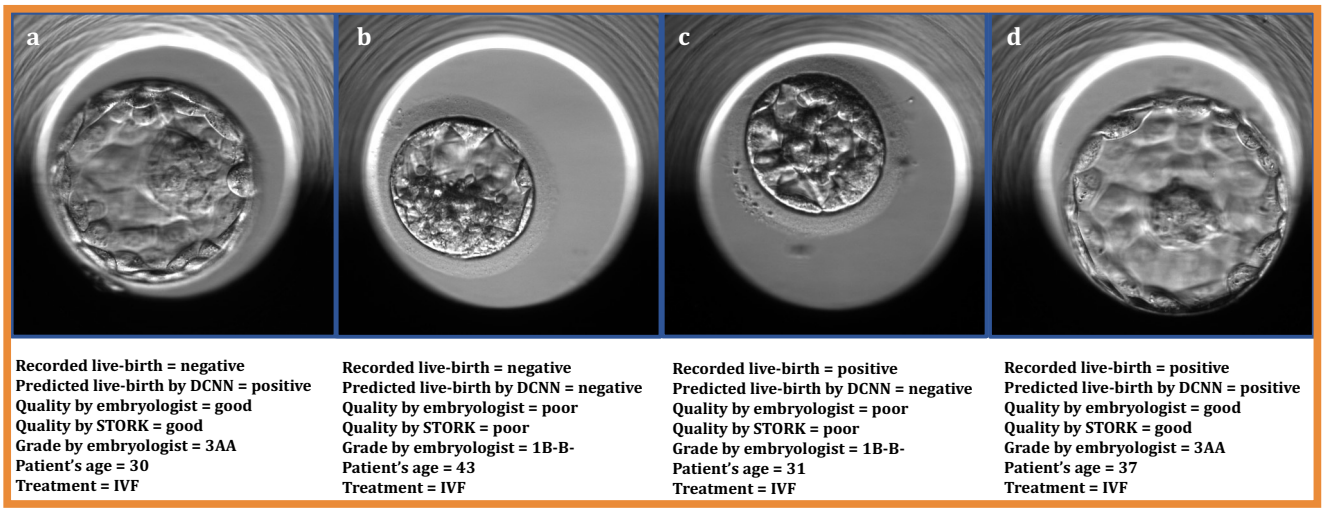


Figure Supplementary 4. The DCNN classifies the embryo images with positive and negative live-birth labels with a focus on their morphological quality. For example, embryos 'a' and 'd' are recorded to negative live-birth and positive live-birth by the laboratory data manager, respectively. DCNN, however, predicted positive live-birth for embryos 'a' and 'd' because they both have good morphological quality. Embryos 'b' and 'c' are recorded as negative live-birth and positive live-birth, respectively. However, the algorithm again classified both embryos 'b' and 'c' as negative live-birth because they have poor-quality.

Figure Supplementary 5

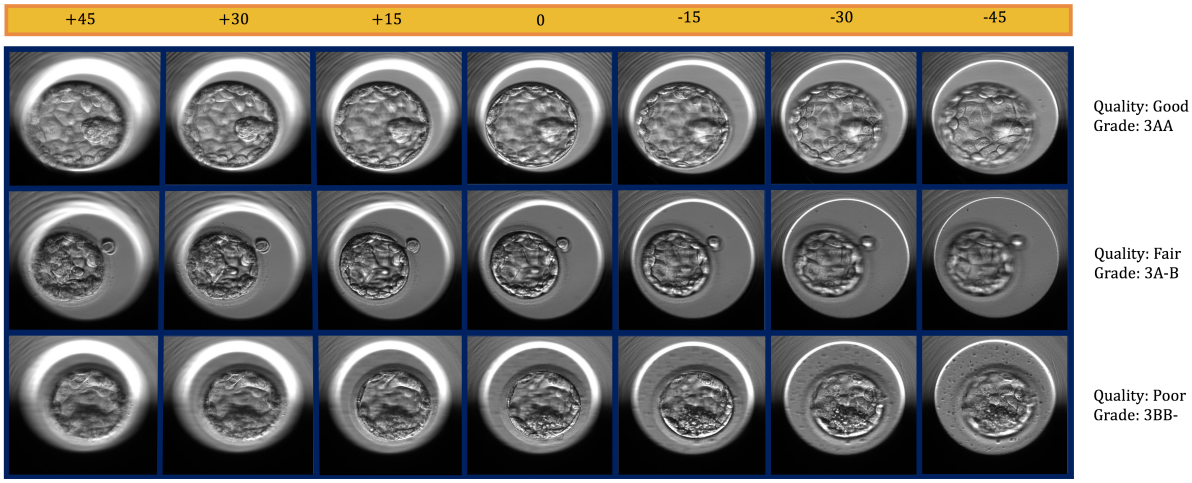


Figure Supplementary 5. This figure shows three examples of Veeck and Zaninovich grades and their corresponding quality labels across seven focal depths.