Original Article

An Artificial Intelligence-Based Algorithm for Predicting Pregnancy Success Using Static Images Captured by Optical Light Microscopy during Intracytoplasmic Sperm Injection

Jared Geller, Ineabelle Collazo[1], Raghav Pai, Nicholas Hendon[1], Soum D. Lokeshwar[2], Himanshu Arora, Manuel Molina, Ranjith Ramasamy

Department of Urology, Miller School of Medicine, University of Miami, [1]Department of Urology, University of Miami – Miller School of Medicine, Miami, FL, [2]Department of Urology, Yale University School of Medicine, New Haven, CT, USA

Abstract

Context (Background): Analysis of embryos for *in vitro* fertilization (IVF) involves manual grading of human embryos through light microscopy. Recent research shows that artificial intelligence techniques applied to time lapse embryo images can successfully ascertain embryo quality. However, laboratories often capture static images and cannot apply this research in a real-world setting. Further, current models do not predict the outcome of pregnancy. **Aims:** To create and assess a convolutional neural network to predict embryo quality using static images from a limited dataset. We considered two classification problems: predicting whether an embryo will lead to a pregnancy or not and predicting the outcome of that pregnancy. **Settings and Design:** We utilized transfer learning techniques using a pretrained Inception V1 network. All models were built using the Tensorflow software package. **Methods:** We utilized a total of 361 randomly sampled static images collected from four South Florida IVF clinics. Data were collected between 2016 and 2019. **Statistical Analysis Used:** We utilized deep-learning techniques, including data augmentation to reduce model variance and transfer learning to bolster our limited dataset. We used a standard train/validation/test dataset split to avoid model overfitting. **Results:** Our algorithm achieved 0.657 area under the curve for predicting pregnancy versus nonpregnancy. However, our model was unable to meaningfully predict whether a pregnancy led a to live birth. **Conclusions:** Despite the limited dataset that achieved somewhat of a lower accuracy than conventional embryo selection, this is the first study that has successfully made IVF predictions from static images alone. Future availability of more data may allow for prospective validation and further generalisability of results.

Keywords: *Deep learning, in vitro fertilization, inception V1, transfer learning*

Introduction

The prevalence of infertility worldwide is about 186 million people and continues to be a large societal burden.[1] Recently, due to economic and societal uncertainty, there has been an even greater psychological impact on infertility patients.[2] Thus, it has become ever more important to improve the efficacy of infertility treatments. The most common assisted reproductive technology used to treat infertility is *in vitro* fertilization (IVF), which developed into an efficacious treatment modality for both male and female infertility, reaching peak live birth rates in 2001–2002.[3] Nevertheless, in the past several years, there has been a decrease in the live birth rates from IVF leading to poorer patient satisfaction and increased

Received: 01-04-2021 **Revised:** 11-07-2021
Accepted: 16-07-2021 **Published:** 28-09-2021

Access this article online
Quick Response Code:
Website:
www.jhrsonline.org
DOI:
10.4103/jhrs.jhrs_53_21

Address for correspondence: Dr. Ranjith Ramasamy,
Department of Urology, University of Miami Miller School of
Medicine, 1120 NW 14th Street, Room 1560, Miami, FL 33136,
USA.
E-mail: ramasamy@miami.edu

How to cite this article: Geller J, Collazo I, Pai R, Hendon N, Lokeshwar SD, Arora H, *et al.* An artificial intelligence-based algorithm for predicting pregnancy success using static images captured by optical light microscopy during intracytoplasmic sperm injection. J Hum Reprod Sci 2021;14:288-92.

costs.[4] While the causes behind this phenomenon are likely multifactorial (including advanced age), the impact of the process of conventional embryo evaluation must be acknowledged.

The current methods of embryo selection for transfer in IVF cycles use morphological evaluations of embryos by trained embryologists. This is done through either conventional microscopic analysis or time-lapse imaging systems.[4] Although the most common scale used for embryo grading is the Gardner scale, which takes into account morphology, rates of cleavage, as well as nutrient uptake and utilization, there is no true universal system that is used.[5] Thus, the traditional method of embryo grading has been regarded to be quite subjective and time-consuming with limited inter-observer and intra-observer agreement among embryologists.[6] When selecting blastocysts for biopsies or cryopreservation, trained embryologists have had average consistency rates of under 60%.[7] This brings into question what future technologies could improve blastocyst selection and ultimately improve live-birth rates. Furthermore, uncertainty often leads to transfer of multiple embryos leading to undesirable multiple pregnancies, which occur at a rate of up to 30% in select settings. Overall, a less subjective process can allow for greater success rates with elective single-embryo transfers, increasing both the efficacy and patient satisfaction associated with IVF.[8]

Artificial intelligence (AI) is an emerging topic in many fields of medicine and has been particularly useful in medical imaging. For example, machine learning algorithms recently have been used to successfully diagnose fractures by analysing skeletal radiographs with accuracy rates comparable to that of experienced orthopedic surgeons.[9] Accordingly, the most recent NIH Roadmap for Foundational Research on AI in medical imaging emphasises the major impact AI will have on addressing medical imaging problems in the next decade.[10] In spite of this, there have only been a few groups that have investigated the use of AI in embryo selection and how they can improve outcomes and patient satisfaction during IVF. As such, the purpose of this study is to evaluate a novel deep learning pipeline that can predict live birth rates using static images from a multi-clinic IVF practice.

Data

The purpose of the AI software was to build an end-to-end model that took in day 5 transferred blastocysts and outputted a variable representing whether the image taken eventually led to a live birth. Following Institutional Review Board Approval (#20201319) and in accordance with the Helsinki Declaration of 1975, data were collected from four clinics in South Florida between the years of 2016 and 2019 and included day 5 photographs of blastocysts taken by optical light microscopy along with other patient features, including SART NUM (embryo identification number), ART PROC (the clinic location), PREG (binary of whether the patient became pregnant), TOTAL ET D5 (number of embryos surviving on day 5), FHB (presence of a fetal heart beat), SACS (the number of gestational sacs on the embryo), DEL (binary of whether the patient delivered a live baby), DEL DATE DONOR AGE, EMBRYO GRADE (quality of embryo), and RECIP STIM START (the date the patient started IVF). These were harvested and stored on a database to aid in evaluation.

The static images, including both high quality [Figure 1] and low quality [Figure 2] embryos, which were each from a unique patient, were taken scaled to a consistent resolution of 224 × 224 pixels, as required by the Inception neural network used by us. The blastocysts were rated by embryologists using a modified version of the Gardner scale, which is analogous to the morphological assessment that is typically used in IVF clinics. Finally, the photos of the blastocysts were distributed into one of three classes: Pregnant and live birth, pregnant and no live birth, and not pregnant.

The available data were augmented to develop a larger set by slightly perturbing images repeatedly and re-adding them to the training set. To augment the data, images were first hand cropped to only show the embryo and exclude excess cell wells. Next, images were scaled to a standardized resolution of 224 × 224. Randomized cropping, rotation, color hue augmentation, scaling, and other data augmentation techniques were utilized to increase the dataset by a factor of 100. With this new, larger dataset, standard deep learning techniques were utilized in the analysis and ultimately two models were run: (i) pregnancy versus nonpregnancy classification (with 207 pregnant 154 not-pregnant images), and (ii) pregnant- and-birth, pregnant- and-no-birth, and not-pregnant 3-class classification (181, 26, and 154 images respectively). To address class imbalance, we utilized resampling techniques.

Methods

STORK, a deep learning model based off of Google's Inception V1, was used as the baseline model.[11] This model takes 224 × 224 sized images as inputs, and outputs a binary Class of 0 or 1 (for pregnant or not pregnant). This model has 27 layers and over five million trainable parameters. STORK's pretrained weights were used, which themselves were trained off of ImageNet. The last layer was then retrained using the data to get the final first model. The second model was created by modifying the last layer of the first model to be a SoftMax with three outputs, one for each of the classes (pregnant and live birth, pregnant and no live birth, and not pregnant).

For both models, 70% of the data was used for training, 20% for validation, and 10% for testing. The final model tuning parameters were as follows: training for 20,000 epochs, a learning rate alpha of 0.012, a weight decay parameter of 0.000005, a batch size of 16, a dropout rate of 20%, and an Adams optimizer gradient decent algorithm. Dropout was computed on the train and validation sets, but not on the test set, and the neural network was randomly initialised using he initialisation.

All result metrics were computed on the test set, which was not touched before the analysis. The researchers only had access to the training and validation sets prior to the single run of the test set. Our AI-based algorithm used pattern recognition from the training set to conclude the efficacy of the blastocyst in its ability to implant.

All statistical methods followed the implementation given in the Tensorflow Slim python library. Beyond this, the receiver operating characteristic curve (ROC) with subsequent calculation of the area under the curve (AUC) was used as the primary statistic to target.

Results

A total of 361 static images were utilized in the algorithm. The model resulted in an area under the curve (AUC) of 0.657, which ranges from 0 to 1 with 1 representing perfect classification. This suggests that the two-class model was able to classify embryos 6% better than the baseline human classification accuracy of 60%.

Figure 3 displays the receiver operating characteristic curve for this model. This shows the true positive rate versus the false-positive rate at various threshold settings. The Receiver operative curve suggests that the model had power to discriminate between the difference classes, and likely would have performed better with a larger dataset.

There was a dip in performance in the three-class model compared to the two-class model. Forty-two percent of live births and 76.4% of nonpregnancies were accurately predicted. Notably, all images in the pregnant and no live birth category were misclassified. This is likely due to under sampling of this class of blastocysts in the data, however, and not to the strength of the model. It simply does not have enough images of this type to learn from. We discuss this limitation further in the discussion section. However, 85.7% of such images were classified

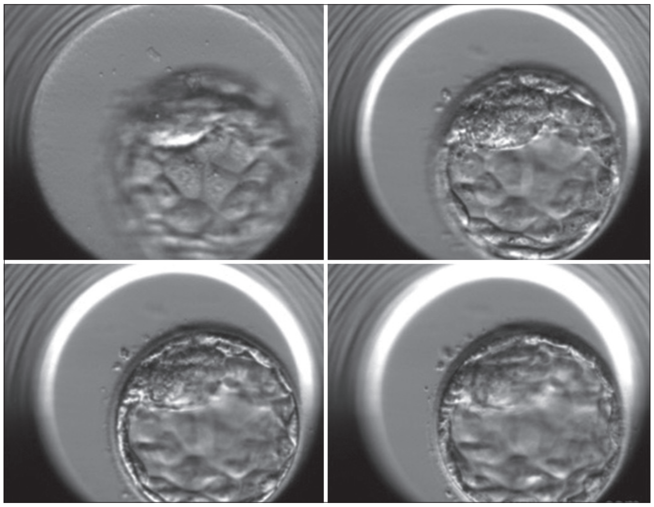

Figure 1: Examples of training images of high quality embryos. The upper right embryo led to pregnancy and no live birth, while the other three led to live births

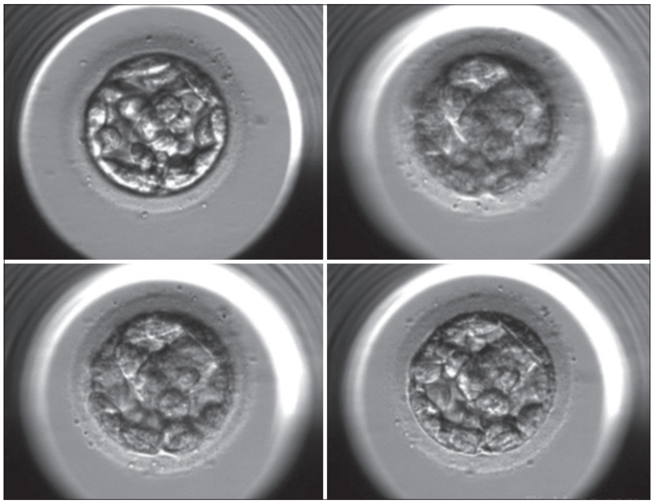

Figure 2: Examples of training images of low quality embryos that did not lead to pregnancy

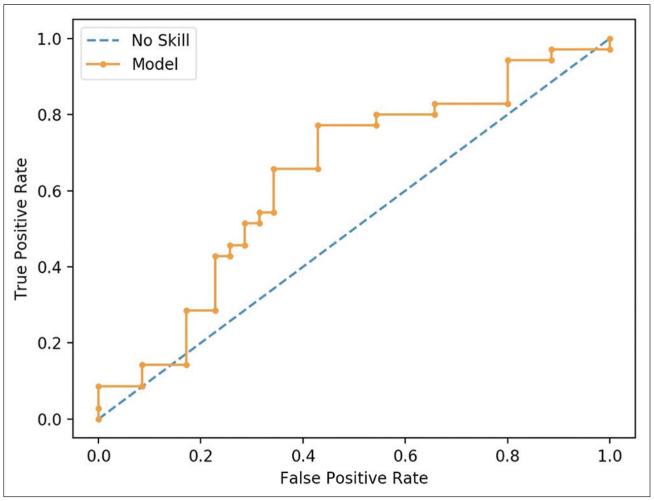

Figure 3: Receiver operating characteristic curve for the artificial intelligence-based model

as not pregnant rather than live birth. This suggests that with more and better data, the model should be able to have predictive power to learn this decision boundary.

Due to the results in the second test, a grad-cam visualization technique on the final layer was performed to visualize what the network was learning to detect. It seemed that grad-cam was looking at the boundary of the blastocyst, as is expected.

Discussion

The majority of suitable deep learning algorithms applied to IVF have focused on capturing embryo images from time-lapse incubators, which allows for continuous monitoring of embryo growth process over time.[12,13] However, many clinics rely on static images and do not have access to time-lapse data. Moreover, the use of time-lapse incubators for embryo grading adds a significant extra cost to IVF treatment. Our hypothesis was that similar accuracy metrics could be achieved with only a small, static image dataset combined with new data augmentation techniques. In a world where "big data" is a hot topic, this research explores the complementary "small data" question applied to IVF. By using image augmentation, dataset balancing, and other techniques, we were able to achieve similar accuracy results to a model with a much larger, feature rich dataset. Beyond having fewer images, our dataset images also had a lower quality, and came from a slightly different data distribution than when compared to those used in STORK. Thus, the main contribution of this study is achieving a suitable classification accuracy for the viability of blastocysts via a static image despite having this limited dataset. In addition, many of the other groups that have attempted to use an AI-based algorithm to predict embryo viability have only measured clinical pregnancy as an endpoint, rather than live birth.[14]

Overall, the rapid advances in AI and its subsequent increase in use across many disciplines of medical imaging is continuing the evolve. Embryo selection, among the most subjective and difficult parts of IVF, is one area in particular where the use of AI could improve outcomes, reduce costs, and standardize processes. While the use of AI in embryo evaluation has been somewhat limited there have been several groups that have only recently begun to test the efficacy of various deep learning pipelines in embryo classification. The first and largest of the embryo selection AI-based studies by Khosravi *et al*. used a deep-learning algorithm that classified 10,148 embryos into either "poor-quality" or "good-quality" at accuracy of 96.64%.[13] Another study by Dirvanauskas *et al*. created an algorithm that can predict embryo development stage with 97.62% accuracy. This is important as the time spent in certain cell-stages through embryo development often correlates with the quality of the embryo and the overall likelihood of pregnancy.[12] This research fits in with the aforementioned studies by (i) validating their results, (ii) showing that a signal still exists despite a smaller dataset, and (iii) generalizing previous models to predict live births.

Although our three-class model did not have enough predicative power to accurately discriminate between the classes: Pregnant and live birth, pregnant and no live birth, and not pregnant, it is clear that with a larger dataset such a problem will be feasible. We would benefit from a dataset with more class balance, and in particular, more images from the underrepresented classes. Moreover, a larger dataset can be incorporated to build a model without utilising transfer learning off of STORK. While this study was able to successfully use static images, which are commonly used in IVF centers, to make clinically relevant predictions, there are still several limitations to address. The samples were all collected from only four different IVF clinics, and the overall sample size was fairly modest at 361 images. This means the model could be open to bias problems, requiring more diverse data. In addition, the blastocysts representing future spontaneous abortion were significantly underrepresented and the images were of poorer quality than many other studies. Despite these limitations, however, the pipeline still allowed for a 0.657 AUC which, while lower than some studies, is reasonable given the sample size and image quality. As such, as more images are incorporated into this study's model, a prospective validation of the pipeline using a larger multi-center database of higher image quality can be performed.

Our design is unique in both the data utilised and the end outcomes measured. This study is among the few that attempted to directly correlate static images to a real clinical endpoint. Furthermore, there is little high-quality evidence showing differences in live birth, miscarriage, stillbirth, or clinical pregnancy in those undergoing IVF with the use of a time-lapse system versus conventional incubation.[15] The majority of other groups measured surrogate outcomes by classifying embryos by quality or other factors that are only associated with increased chances of successful pregnancy and live birth. This study, therefore, serves as a validation of the STORK model, while also leaving open the possibility of improvement through newer models, larger datasets, and targeting a broader multi-class classification problem.

Conclusions

The use of AI in clinical medicine is continuing to evolve. Namely, the use of AI in embryo selection for IVF has the potential to reduce subjectivity and thereby improve outcomes and patient satisfaction. This study is among the first that demonstrated how AI algorithms could be used to predict pregnancy success by using static images that are so commonly used in IVF clinics worldwide. The availability of more data will allow for prospectively validation and further generalizability of the results.

Acknowledgments

We would like to acknowledge Noah Jacobson for his valuable insight and advice regarding our use of Transfer Learning and Hyperparameter Tuning.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

Funding: No funding was received for conducting this study.

Availability of data and material: Not applicable.

Code availability: The code used for the algorithm is available at the following web-site: https://github.com/jgeller1/IVFFinal/

References

1. Wasilewski T, Łukaszewicz-Zając M, Wasilewska J, Mroczko B. Biochemistry of infertility. Clin Chim Acta 2020;508:185-90.
2. Vaughan DA, Shah JS, Penzias AS, Domar AD, Toth TL. Infertility remains a top stressor despite the COVID-19 pandemic. Reprod Biomed Online 2020;41:425-7.
3. Gleicher N, Kushnir VA, Barad DH. Worldwide decline of IVF birth rates and its probable causes. Hum Reprod Open 2019;2019:hoz017.
4. Abreu CM, Thomas V, Knaggs P, Bunkheila A, Cruz A, Teixeira SR, *et al*. Non-invasive molecular assessment of human embryo development and implantation potential. Biosens Bioelectron 2020;157:112144.
5. Gardner DK, Sakkas D. Assessment of embryo viability: The ability to select a single embryo for transfer-a review. Placenta 2003;24 Suppl B:S5-12.
6. Storr A, Venetis CA, Cooke S, Kilani S, Ledger W. Inter-observer and intra-observer agreement between embryologists during selection of a single day 5 embryo for transfer: A multicenter study. Hum Reprod 2017;32:307-14.
7. Bormann CL, Thirumalaraju P, Kanakasabapathy MK, Kandula H, Souter I, Dimitriadis I, *et al*. Consistency and objectivity of automated embryo assessments using deep neural networks. Fertil Steril 2020;113:781-7.e1.
8. Muglia VF, Prando A. Renal cell carcinoma: Histological classification and correlation with imaging findings. Radiol Bras 2015;48:166-74.
9. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, *et al*. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017;88:581-6.
10. Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, *et al*. A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 NIH/RSNA/ACR/the academy workshop. Radiology 2019;291:781-91.
11. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019;6:60.
12. Dirvanauskas D, Maskeliunas R, Raudonis V, Damasevicius R. Embryo development stage prediction algorithm for automated time lapse incubators. Comput Methods Programs Biomed 2019;177:161-74.
13. Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, *et al*. Deep learning enables robust assessment and selection of human blastocysts after *in vitro* fertilization. NPJ Digit Med 2019;2:21.
14. VerMilyea M, Hall JM, Diakiw SM, Johnston A, Nguyen T, Perugini D, *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;35:770-84.
15. Armstrong S, Bhide P, Jordan V, Pacey A, Marjoribanks J, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. Cochrane Database Syst Rev 2019;5:CD011320.