Human Reproduction, pp. 1–2, 2022
https://doi.org/10.1093/humrep/deac131

human reproduction SUPPLEMENTARY DATA

Supplementary materials and methods

Image augmentations used during training

During the training process, additional image augmentations were applied (including rotations, reflections, Gaussian blur, contrast variation, horizontal and vertical translations jitter), and random compression or jpeg noise were included as described in Zheng *et al.* (2016). In addition, new methods were used for augmentations including color rotation (random swapping of bias among red, green and blue channels of images during training), border additions (random addition of image borders so the artificial intelligence (AI) learns to ignore those features in incorrectly cropped images), resolution (randomly selected resolutions for each batch, so that the training is robust to the input resolution of the image) and coarse drop-out (introduction of random image artifacts, such as limited-sized regions of black pixels so that the training is robust to image noise artifacts).

Details regarding the model training and selection process

The first step in the model selection process involved applying the untrainable data cleansing (UDC) method (Dakka *et al.*, 2021) and cleansed training datasets were obtained for a range of UDC thresholds and three segmentation styles: no segmentation, zona segmentation and intra-zonal cavity (IZC) segmentation. IZC segmentation indicates that the training dataset was curated through model configurations that require input images that masked the *zona pellucida* and background region of the image (as black), exposing the IZC and inner cell mass regions. Zona segmentation indicates that model configurations require inputs of images masking the IZC region. Zona segmentation models were considered during the model selection process, but the final model described below did not include any zona models. The UDC method represents a unique contribution which is not typically employed in AI training, and the computer-vision preprocessing methods to isolate the IZC and zona regions of the embryo image prior to deep learning were developed in VerMilyea *et al.* (2020).

The neural network architectures considered during training represent standard architectures and were updated only to include a binary classifier layer and softmax layer. A range of architectures was considered, with the final chosen solely on performance on the holdback validation dataset. Hyperparameters were also chosen with a diverse range of options and further refined during training to optimize each value.

To measure the performance of the trained models to guide selection, a holdback validation dataset was chosen from the union of the source images among the candidate training datasets (n = 300), which was kept identical for all model configurations and segmentation styles, to act as a benchmark for comparison. The shared holdback validation dataset was used to select teacher models for distillation, then candidate models for inclusion in an ensemble model, and finally for selecting the final ensemble model. While distillation has been previously described in the field of machine learning (Hinton *et al.*, 2014), the number of models chosen to distill together is treated as an additional hyperparameter to optimize during training. In this study, it was found that between 1 and 3 models were optimal.

The deep learning optimization specifications were the same as those described in VerMilyea *et al.* (2020). In brief, each deep neural network used weight parameters obtained from pretraining on ImageNet, with the final classifier layer replaced with a binary classifier corresponding to aneuploid and euploid classification. Training of AI models was conducting using PyTorch library (version 1.3.1 including Torchvision version 0.4.2; Adam Paszke, Sam Gross, Soumith Chintala and Gregory Chanan; 1601 Willow Rd, Menlo Park, CA 94025, USA), with CUDA support (version 9; Nvidia Corporation; 2788 San Tomas Expy, Santa Clara, CA 95051, USA). The optimizations considered are standard in the industry, with novel methods applied in preparation of the training dataset, including UDC and removal of mosaic embryos prior to training to reduce label noise in the dataset.

The following methods were used to evaluate constituent models to select the final ensemble model:

- *Confidence metrics for translatability:* Multiple confidence metrics were defined, which were considered more robust indicators of translatability (ability of the model to generalize to unseen data) than accuracy measures. The two confidence metrics used in the final model selection were log loss and tangent score.
- *Model stabilization:* The stability of the selection metric value on the validation dataset was measured over all epochs of the training process.
- *Prediction accuracy:* Which models provided the best validation accuracy, for both classes individually: euploid and aneuploid labeled embryos, the total combined accuracy and the balanced accuracy (defined as the weighted average accuracy across both class types of embryos) was recorded, and the result of accuracy metrics on the test dataset were used to determine if the final model selected had translated well. AUC/receiver-operating characteristic and AUC/precision-recall curves were also evaluated. In all cases, use of ImageNet pretrained weights demonstrated improved performance of these quantities.

The process of selecting models on a validation dataset prior to reporting on a test dataset is standard practice, but the reliance on a confidence-based metric rather than an accuracy-based metric, and the use of tangent score as a selection metric for classes that include a greater degree of noise, were novel contributions in this study.

The first round of training identified a set of potential teacher models, which were selected on the aforementioned set based on the metric tangent score. One to three teachers were selected for both segmentation styles, and a second round of models was trained. Both student and teacher models were considered (from both training rounds) as candidates for ensembling. Ensembling was performed as described in Maclin and Opitz (2011), with voting strategies evaluated including mean, median, max and majority-mean voting. The use of an ensemble method is standard practice in machine learning, with the choice of voting strategies employed treated as hyperparameters and only optimized on the validation dataset.

Methods to address overfitting

Training and selection of constituent models making up the final ensemble AI model were performed primarily using two confidence score metrics, log loss and tangent score. These were prioritized over traditional accuracy metrics to select models that would be more likely to generalize well to unseen datasets. The dependence of log loss on the confidence score assigned to an image during training is highest for confidently misclassified images. The dependence of tangent score on the confidence score for an image is equally high for confidently misclassified and correctly classified images. As a result, the two confidence metrics are highly complementary when dealing with levels of noise distributed unequally among the classes, which in this case are euploid and aneuploid embryos, as both of these metrics take into account the distribution of the AI scores (i.e. to what extent they have correctly or incorrectly classified an embryo). This differs from the use of accuracy as a selection metric, where a poorly performing model can overfit by chance, with a score distribution akin to a lop-sided Gaussian, where only a slight change in scores and distribution can significantly change the accuracy. Instead, the clusters of scores are well-separated when optimizing on confidence score metrics, leading to more robust model selection for translatability.

Overfitting of the genetics AI algorithm in the current study was tested for by comparing the performance of the final ensemble model to that of the individual constituent AI models making up the final model. It was found that the three constituent models had overall accuracies of 65.4%, 62.5% and 65.5% on the uncleansed Day 5 blind test dataset, similar to that of the final ensemble model (65.3%). The log loss values, representing one of the primary metrics used to select constituent models, were 0.78, 0.80 and 0.84 for the constituent models, compared to the superior value of 0.75 for the final AI model. These performance values demonstrate that the ensemble has not been overfitted to the validation dataset during training.

References

Dakka MA, Nguyen TV, Hall JMM, Diakiw SM, VerMilyea M, Linke R, Perugini M, Perugini D. Automated detection of poor-quality data: case studies in healthcare. *Sci Rep* 2021;**11**:18005.

Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. In: *Neural Information Systems (NIPS) Deep Learning Workshop*. 2014. https://arxiv.org/abs/1503.02531 (12 December 2014, date last accessed).

Maclin R, Opitz D. Popular ensemble methods: an empirical study. *IEEE Comput Soc Conf Comput Vis Pattern Recogn* 2011: 4480–4488.

VerMilyea M, Hall JMM, Diakiw SM, Johnston A, Nguyen T, Perugini D, Miller A, Picou A, Murphy AP, Perugini 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**:770–784.

Zheng S, Song Y, Leung T, Goodfellow I. Improving the robustness of deep neural networks via stability training. *IEEE Comput Soc Conf Comput Vis Pattern Recogn* 2016:4480–4488.